NOTE TO USERS

Performance Improvement of Mobile IP-Based Handoff Between WLAN and GPRS Networks

Mahmud Hossain

A Thesis

In

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Science at

Concordia University

Montreal, Quebec, Canada

May 2004

Your file Votre référence
ISBN: 0-612-94699-1
Our file Notre référence
ISBN: 0-612-94699-1

Abstract

Performance Improvement of Mobile IP-Based Handoff Between WLAN and GPRS Networks

Mahmud Hossain

Next generation wireless networks are characterized by heterogeneity particularly in terms of the underlying technology. However, the end users would like to enjoy seamless mobility anywhere in the network while using higher bandwidth at a lower cost. Wireless Local Area Networks (WLANs) have emerged as a new technology with a promise of very high bandwidth, which can compliment the widely deployed General Packet Radio Service (GPRS) networks. However, one of the challenges of the heterogeneous networking of WLAN and GPRS is to manage the handoff, particularly for the voice (real time) services.

This thesis presents a handoff technique between the IEEE 802.11 WLAN, which covers a small area, and the GPRS cellular network that overlays the WLAN and covers a larger area. Mobile IP is chosen for managing the handoff to accommodate the all-IP vision of the future interconnected networks. However, the handoff management, as proposed in Mobile IP, is mainly for the data services. Therefore, it would be a considerable challenge to achieve low latency handoff for voice services, particularly for the lost packet recovery mechanism as proposed in Mobile IP. While Multicasting can be adopted to reduce latency, it is associated with higher overhead as the user increases its speed.

In this thesis, a new handoff technique is proposed. The technique reduces the handoff delay, recovers the lost packets during the handoff transition period with improved overhead, increases the throughput, and also prevents unnecessary handoffs from due to ping-pong effect. It includes an algorithm to be implemented in the handoff decision making process as well as modification of the Mobile IP for handoff signaling management to achieve improved performance in latency, overhead, and throughput.

Acknowledgement

First and foremost I would like to extend my deep and sincere gratitude to my respected supervisor Dr. Ahmed K. Elhakeem. His wide knowledge and his logical way of thinking have been of great value for me. He provided a motivating, enthusiastic, and critical atmosphere during the many discussions we had. His understanding, encouraging and personal guidance have provided a good basis for the thesis.

I sincerely acknowledge Dr. Walaa Hamouda who, as my second supervisor, provided constructive comments during my thesis time. His integral view on research and his mission for providing 'only high-quality work and not less', has made a deep impression on me. I owe him lots of gratitude for his stimulating suggestions.

I would like to thank Dr. M. R. Soleymani, Dr. Y. R. Shayan and Dr. K. Siddiqui, who, as the members of the examination committee, have provided necessary suggestions for overall improvement of the thesis work.

It is a pleasure to record my personal debt to my friend Faridul Islam, who helped me a lot during the development of the simulation program. This thesis was enriched significantly through helpful discussions with my friends Shahidul Akond, Ejaj Mahfuz, Rajib Ullah, Maruf Khondker, and Faridul Islam. I owe special thanks to them.

No words can express my debt to my family: my wife Mahbina Hossain, my daughter Bushra Hossain and my son Mujtoba Hossain, who have always put my interest, related to my research works, before theirs. Particularly, I would like to recognize the unrelenting support my wife Mahbina afforded me.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

1.1. Background

Wireless systems have been developed in an evolutionary way by generation over the last two decades or so. First Generation (1G) systems are of diminishing importance. Second Generation (2G) cellular systems, mostly Global System for Mobile Communications (GSM) and cdmaOne, have enabled a high level of mobility with wire equivalent quality for voice and low-speed data services. Although 2G technology is adequate in meeting the voice communication needs for the typical cellular subscribers, its data communication capabilities are cumbersome and limited [1]. In contrast, wired data services, which offer higher bandwidth than those of wireless data services, have grown in popularity due to the availability and the affordability. To compete with the wire services technology, Third Generation (3G) cellular systems promise competitive data rates almost similar to that of wired technology.

Due to the delay of the 3G cellular networks and the large investments made for new spectrum in which to offer 3G services, the cellular industry is now looking for the ways to offer "*3G-like*" services in efforts to generate new revenue stream in today's environment. 2.5G cellular data technology, in particular General Packet Radio Service (GPRS), has gained supports as a wide area data solution [15] [43]. However, due to

limitations in the data rates, the use of GPRS systems is restricted in business and multimedia applications [1].

On the other hand, Wireless Local Area Network (WLAN), particularly IEEE 802.11, has become one of today's driving wireless technologies, delivering high data rates on unlicensed spectrum for both enterprise and the home [10] [12]. Besides transfer of purely data traffic, WLANs can also support packetized voice transmission [2]. The infrastructure of WLAN costs the end-user much less than that of the infrastructure of the cellular phones. Moreover, WLAN radio technology provides superior bandwidth compared to that of the current 3G cellular technology.

Despite recent advances in the achieved bandwidth of all types of wireless networking technology, indoor networks continue to provide much higher bandwidth than outdoor networks. The bandwidth gap is expected to either persist or widen in the near future [3]. Therefore, a natural choice for the wireless industry is to integrate WLANs and cellular networks to provide a seamless user experience. Bridging WLANs and cellular networks will however require an interworking mechanism capable of providing integrated authentication, integrated billing, roaming, terminal mobility, and service mobility [1].

Most existing wireless network technologies can be divided into two categories: i) those that provide a low-bandwidth service over a wide geographic area (e.g. GPRS) and ii) those that provide a high bandwidth service over a narrow geographic area (e.g. WLAN). While it would be desirable to provide a high-bandwidth service to mobile users at all

times, this is unlikely. WLANs only provide limited coverage, and a mobile host equipped only with a wide-area network interface cannot exploit existing high-bandwidth infrastructure, such as in-building wireless local area networks or wired networks. The solution is to use a combination of wireless networks, defined as *wireless overlay network*, to provide the best possible coverage over a range of geographic areas [4]. The integrated network of WLANs and GPRS is a typical example of *wireless overlay network* as well as a part of the next generation networks, which would be featured by a heterogeneous communication environment [4].

If the mobile terminal wants to roam between different networks when it is in communication, the handoff mechanism between the heterogeneous networks become a critical issue [5]. A heterogeneous (or hybrid) network can be defined as a network which comprises of two or more different (not homogeneous) technology to provide ubiquitous coverage [39]. There has been some research works in recent time to address the issue of handoff between heterogeneous networks [1] [3] [4] [5] [6] [41] [43] [44] [45].

Heterogeneity presents the major challenge for mobility management. From a protocol stack perspective, the Network Layer is the lowest possible layer where convergence of the heterogeneous wireless systems can be developed since the major difference between them lies in the two lower layers (i.e. Physical Layer and Data Link Layer). Besides, the desire to extend the great success of the Internet Protocol (IP) in the wired world to wireless leads to an all-IP vision [6]. IP has so far been the best integration technology for heterogeneous networks, and there is currently no foreseeable alternative of IP [7]. To

allow seamless handoff to take place in IP-based heterogeneous networks, IP must support users' mobility. Internet Engineering Task Force (IETF) has developed a standard, namely Mobile IP, to support the mobility in IP. Mobile IP provides a network layer solution to node mobility across the IP network [8]. We chose Mobile IP for managing Handoff (HO) between WLAN and GPRS.

1.2. Objectives of the Thesis

The main objectives of the thesis are as follows:

1. To design an algorithm for handoff decision making process suitable for physical properties of the WLAN and the GPRS networks.
2. To develop a technique for Mobile IP-based handoff between WLAN and GPRS networks which improves various Quality of Service (QoS) criteria (e.g. Latency, Overhead, Throughput etc)
3. To observe and analyze the performance of the proposed handoff technique, and to compare the results with the existing/conventional techniques.

1.3. Organization of the Thesis

The remainder of the thesis is organized as follows:

Chapter 2 provides a brief overview of WLAN, GPRS, and Mobile IP standards. However, it emphasizes on the areas related to mobility management of the above technologies.

Chapter 3 investigates issues related to handoff between WLAN and GPRS networks. This chapter discusses the conventional handoff algorithms used in wireless networks,

and different mobility architectures between WLAN and GPRS networks. Finally, it elaborates the handoff latency and possible lost packet recovery mechanisms involve in Mobile IP-based handoff.

Chapter 4 highlights the principal contributions of this thesis. The chapter, at the beginning, describes our proposed handoff scheme with the aid of necessary flow charts to explain the underlying algorithms. It outlines the simulation model as well as the parameters/assumptions adopted in the model. Finally, the chapter extensively illustrates the performances of the four important handoff criteria; handoff, overhead, throughput, and call drop.

Chapter 5 summarizes the results and outcome of the thesis. It also provides some suggestions for future research.

Chapter 2

2. Overview of WLAN, GPRS, and Mobile IP

2.1. Overview of WLANs

The basis for the WLAN technology was developed in World War II by the U.S. Military as a way for securing the safe, private delivery of voice communications without eavesdropping by the enemy [9]. WLANs even predate wired LANs because ALOHA, the basis of Ethernet, was the first LAN and it was radio-based.

2.1.1. WLAN Technology

WLANs are generally categorized according to the transmission technique that is used. All current WLAN technologies fall into one of the following categories:

I) ***Infrared (IR):*** An individual cell of an IR LAN is limited to a single room, because infrared light does not penetrate opaque walls.

II) ***Spread Spectrum:*** This type of LAN makes use of spread spectrum transmission technology.

III) ***Narrowband:*** These LANs operate at narrowband microwave frequencies.

IV) ***Carrier Current:*** This technique uses power lines as a medium for the transport of data.

Currently, the most popular type of WLANs use spread spectrum techniques. Spread spectrum spreads the signal power over a wider band of frequencies, sacrificing

bandwidth to gain signal-to-noise performance. This contradicts the desire to conserve frequency bandwidth, but the spreading process makes the data signal much less susceptible to electrical noise than conventional radio modulation techniques [10]. Figure 2.1 demonstrates the spreading of narrowband signal.

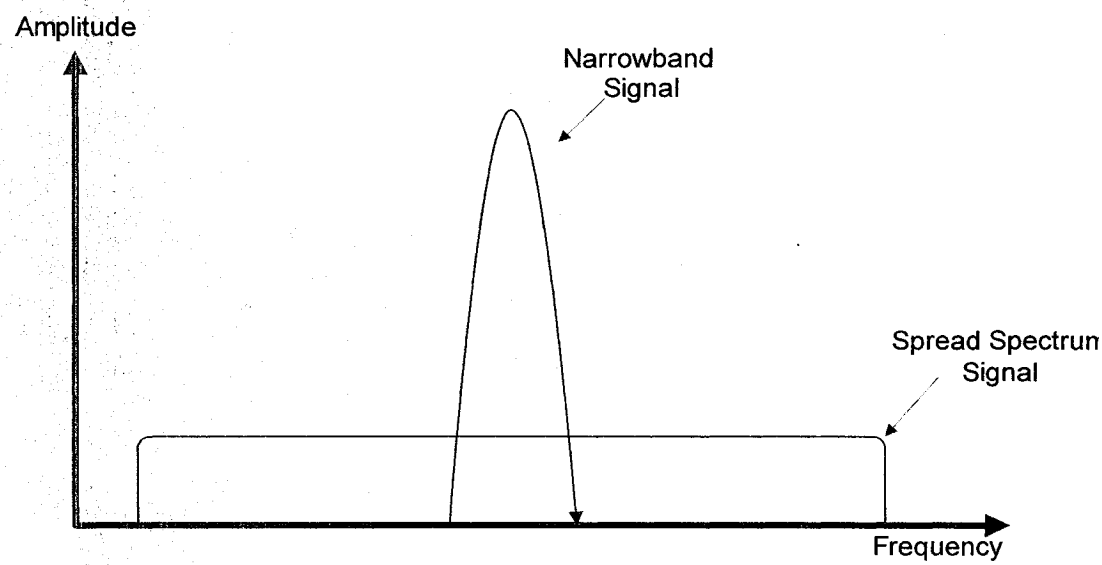

Figure 2.1: Spread Spectrum

2.1.2. Frequency Bands

WLANs mostly operate in unlicensed bands, popularly known as Industrial, Scientific, and Medical (ISM) bands. In 1985, the Federal Communications Commission (FCC) modified the radio spectrum regulations for unlicensed devices, and authorized WLANs to operate in the ISM bands [11], as shown in Figure 2.2.

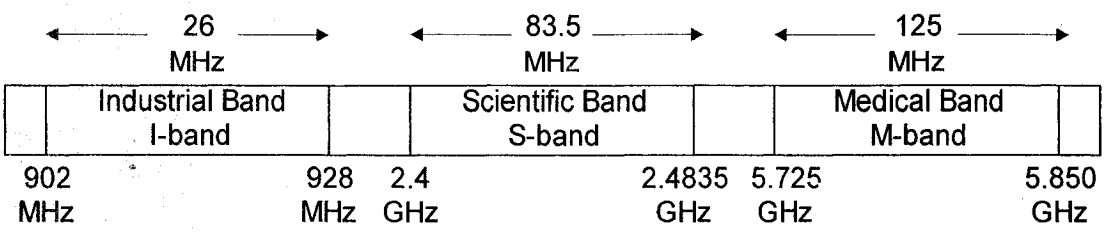

Figure 2.2: ISM Frequency Band

Many WLANs deployed today in the United States operate at 902 MHz, but this frequency is not available throughout the world.

2.1.3. Standardization

The development of WLANs was initially motivated by the allocation of spectrum in the ISM band by the FCC when IEEE 802.4 standard body considered extensions of token bus technology to wireless transmission. This led to the creation of IEEE 802.11 in May 1989. However, the initial standards activity was very contentious, and progress was slow. In October of 1997, the first completed standard from the IEEE 802.11 body was ratified. Since then continuous standardization work is going on in the field of WLANs. The IEEE 802.11 Working Group has been working to add undefined sections through the efforts of several Task Groups (TGs).

In Europe, High Performance Radio Local Area Network (HIPERLAN) Type 2 standard is defined by European Telecommunications Standards Institute (ESTI) in April 2000 [48]. HIPERLAN Type 2 has a centralized Medium Access Control (MAC) as opposed to distributed MAC of IEEE 802.11.

The standard resulted from the standardization process of the IEEE, which is officially called IEEE Standard for Wireless LAN Medium Access (MAC) and Physical Layer (PHY) Specifications, defines over-the-air protocols necessary to support networking in a local area. Our discussion of WLANs is limited to the IEEE 802.11 standard.

2.1.4. IEEE 802 Standards Family and OSI Reference Model

The IEEE 802 family of standards falls within the scope of layers 1 and 2 of the OSI Reference Model. The Logical Link Control (LLC) protocol specifies the mechanisms for addressing stations across the medium and for controlling the exchange of data between two stations; whereas, the MAC and PHY Layers provide medium access and transmission functions. Figure 2.3 compares the IEEE 802 Protocol Layers to OSI Reference Model [10].

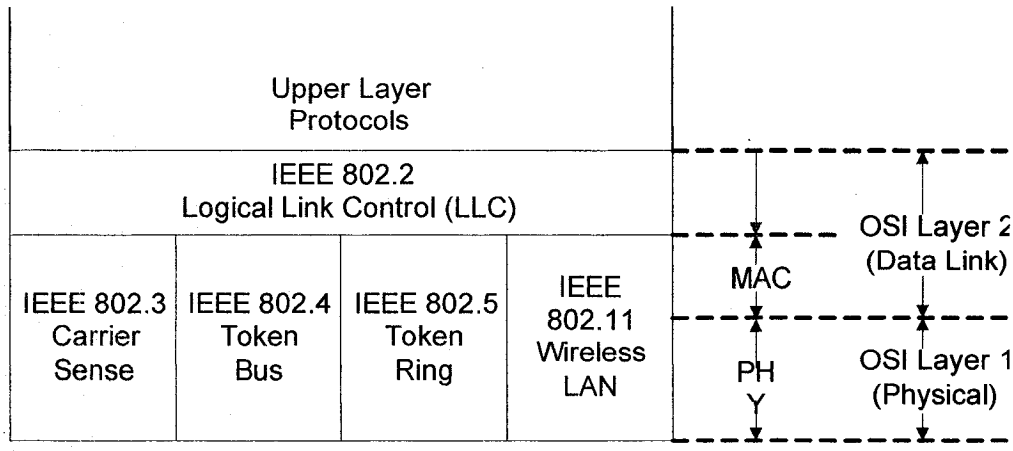

Figure 2.3: IEEE 802 Standards Family and OSI Model

2.1.5. IEEE 802.11 Architecture

2.1.5.1. Physical Topology

The IEEE 802.11 physical topology consists of components, interacting to provide a WLAN that enables station mobility transparent to higher protocol layers, such as the LLC. Figure 2.4 illustrates the model [13] developed by the IEEE 802.11 Working Group.

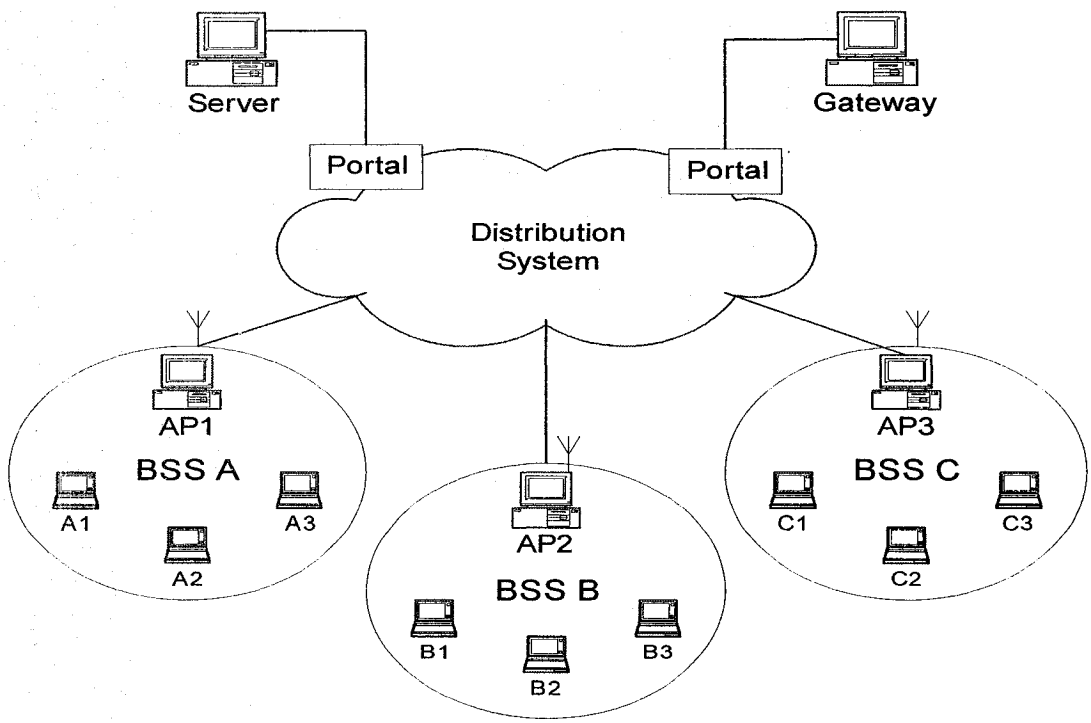

Figure 2.4: IEEE 802.11 Topology

The smallest building block of a WLAN is a Basic Service Set (BSS), which consists of a number of stations executing the same MAC protocol and competing for access to the same shared wireless medium. A single BSS can be used to form an ad hoc network, which is typically temporary in nature. A BSS may be isolated or it may connect to a

backbone Distribution System (DS) through an Access Point (AP). It is also possible for two BSSs to overlap geographically, so that a single station could participate in more than one BSS. Further, the association between a station and a BSS is dynamic.

The AP functions as a bridge. It broadcasts a beacon signal periodically (typically in every 100 ms) [39]. The MAC protocol may be fully distributed or controlled by a central coordination function housed in the AP. The BSS generally corresponds to what is referred to as a cell in cellular telephony literature. The DS can be a switch, a wired network, or a wireless network. A set of BSSs can be interconnected by a DS to form an Extended Service Set (ESS). The ESS appears as a single logical LAN to the LLC level. An ESS can also provide gateway access for WLAN users into a wired network such as the Internet. This access is accomplished via a device known as portal. The term Infrastructure Network is used informally to refer to the combination of BSSs, a DS, and portals [12].

2.1.5.2. Logical Architecture

The logical architecture of 802.11 standard that applies to each station consists of a single MAC Layer and one of the three separate Physical Layers (i.e. Frequency Hopping Spread Spectrum, Direct Sequence Spread Spectrum, and Infrared Light).

2.1.6. IEEE 802 Services

IEEE 802.11 defines nine services that need to be provided by the WLAN to yield functionality equivalent to that which is inherent to wire LANs.

Association: Each station must initially invoke the association service with an AP before it can send information through a DS. Each station can associate with only a single AP, but each AP can associate with multiple stations.

Disassociation: A station or AP may invoke the disassociation service to terminate an existing association. This service is a notification; therefore, neither party may refuse termination. Stations should disassociate when leaving the network.

Reassociation: The reassociation service enables a station to change its current state of association from one AP to another. This keeps the DS informed of the current mapping between AP and station as the station moves from BSS to BSS within an ESS.

Distribution: A station uses the distribution service every time it sends MAC frames across a DS. The distribution service provides the DS with only enough information to determine the proper destination BSS.

Integration: The integration service enables the delivery of MAC frames through a portal between a DS and a non-802.11 LAN. The integration function performs all required media or address space translations.

Authentication: All 802.11 stations, whether they are part of an independent BSS or ESS network, must use the authentication service prior to establishing a connection with another station with which they will communicate. The standard defines two authentication services – (i) open system authentication, (ii) shared key authentication.

Deauthentication: When a station wishes to disassociate with another station, it invokes the deauthentication service. Deauthentication is a notification, and cannot be refused.

Privacy: IEEE 802.11 offers a privacy service option that raises the security level of the 802.11 network to that of a wired network. The privacy service is based on the 802.11 Wired Equivalent Privacy (WEP) algorithms.

2.1.7. Medium Access Control (MAC) Layer

Each station and AP on an 802.11 WLAN implements the MAC Layer service, which provides the capability for peer LLC entities to exchange MAC Service Data Units (MSDUs) between MAC Service Access Points (SAPs). The MSDUs carry LLC-based frames that facilitate functions of the LLC Layer [10]. The MAC Layer provides the following primary operations: (i) accessing the wireless medium, (ii) joining a network, (iii) providing authentication and privacy.

The IEEE 802.11 MAC protocol is specified in terms of coordination functions that determine when a station in a BSS is allowed to transmit and when it may be able to receive PDUs over the wireless medium. The Distributed Coordination Function (DCF) provides support for asynchronous data transfer of MSDUs on a best-effort basis. Under the DCF, the transmission medium operates in the contention mode exclusively, requiring all stations to contend for the channel for each packet transmitted. The standard also defines and optional Point Coordination Function (PCF), which may be implemented by an AP, to support connection-oriented time-bounded transfer of MSDUs. Under PCF the medium can alternate between the Contention Period (CP), during which the medium uses contention mode, and a Contention-free Period (CFP). During the CFP, the medium

usage is controlled by the AP, thereby eliminating the need for stations to contend for channel access [12]. Figure 2.5 depicts the MAC architecture.

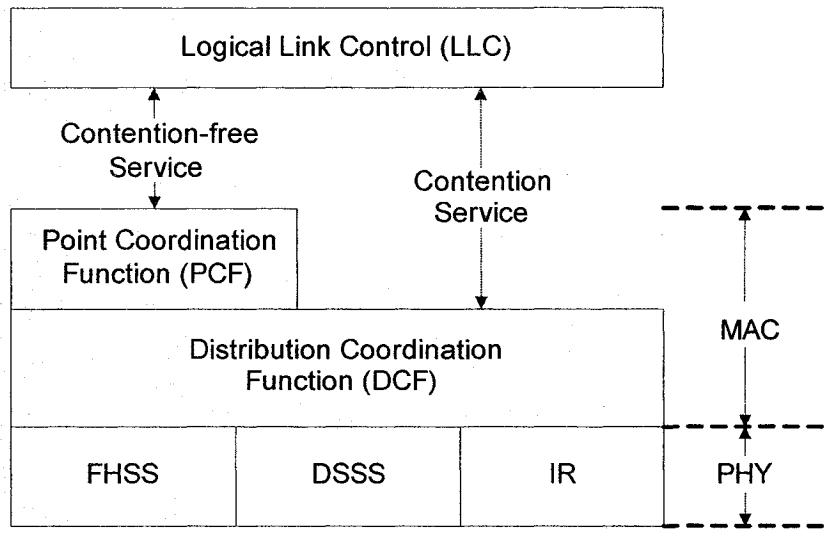

Figure 2.5: IEEE 802 MAC Architecture

2.1.7.1. Distributed Coordination Function (DCF):

The DCF makes use of Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA) protocol. If a station has a MAC frame to transmit, it listens to the medium. If the medium is idle, the station may transmit; otherwise the station must implement a backoff algorithm, and wait until the current transmission is complete before transmitting. To ensure the smooth and fair functioning of this algorithm, DCF includes a set of delays, known as Interframe Spaces (IFSs) that amounts for a priority scheme. To minimize collisions and maximize throughput, the standard defines a contention window that increases exponentially with each retransmission. The flowchart in Figure 2.6 illustrates the operation of DCF.

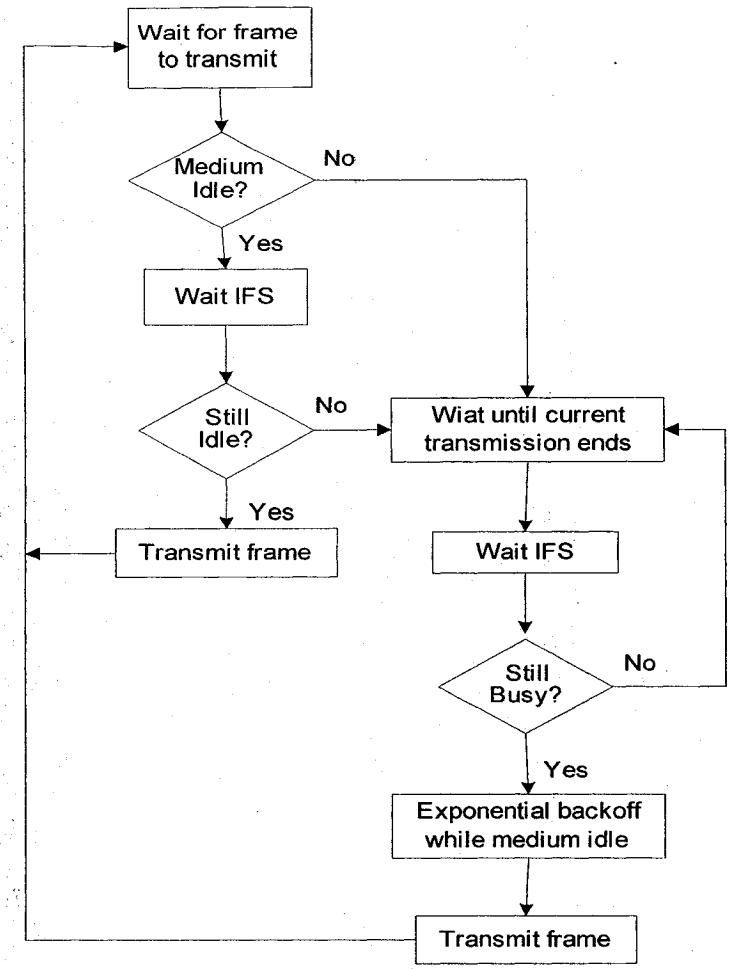

Figure 2.6: Operation of DCF using CSMA-CA

A handshake procedure, using some control frames, is developed to operate with CSMA-CA when there is a hidden station problem. The following describes each of the IFS intervals:

Short IFS (SIFS): The SIFS is the shortest of the IFSs, providing the highest priority level by allowing some frames to access the medium before others.

PCF IFS (PIFS): The PIFS, a midlength IFS, is the interval that stations operating under the PCF use to gain access to the medium.

DCF IFS (DIFS): All stations operating according to the DCF use the DIFS interval, which is the longest IFS, for transmitting data and management frames.

Figure 2.7 shows a time diagram [12] that demonstrates DCF contention.

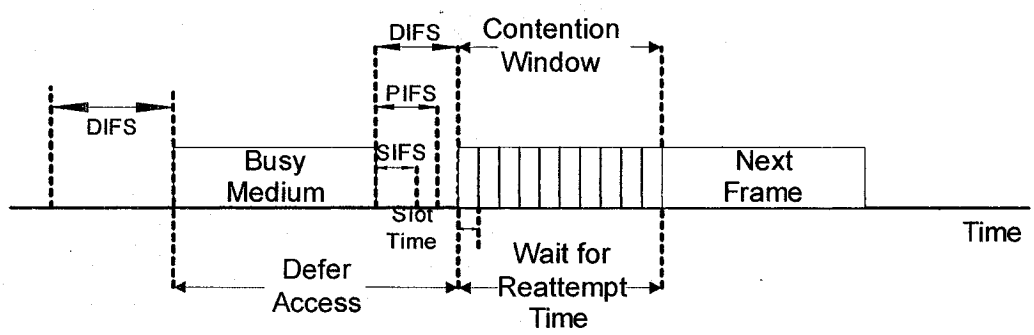

Figure 2.7: Time Diagram of DCF

2.1.7.2. Point Coordination Function (PCF)

The PCF is an optional capability that can be used to provide connection-oriented, contention-free services by enabling polled stations to transmit without contending for the channel. The PCF makes use of PIFS when issuing polls. Because PIFS is smaller than DIFS, the point coordinator can seize the medium and lock out all asynchronous traffic while it issues polls and receive responses.

PCF is built on top of DCF and exploits features of DCF to assure access for its users. The maximum size of CFP is determined by the manageable parameter, CFP_Max_Duration. It is up to the AP to determine how long to operate the CFP during any repetition interval. At the beginning of each CFP repetition interval, all stations in the BSS update their Network Allocation Vector (NAV) to the maximum length of the CFP.

During the CFP, stations may transmit only to respond to a poll from the point coordinator or to transmit an acknowledgement (ACK) one SIFS interval after receipt of an MPDU. Figure 2.8 shows a sketch [12] of the CFP repetition interval.

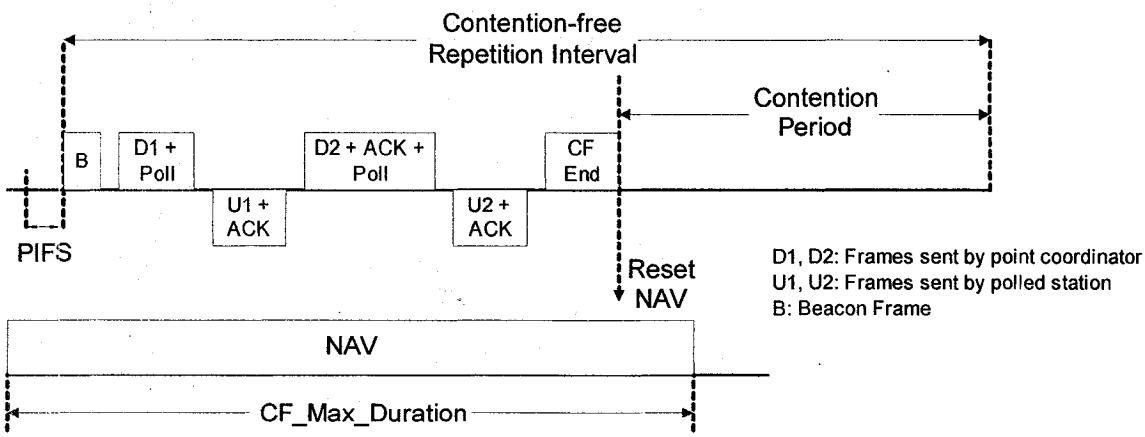

Figure 2.8: Point Coordination Frame Transfer

2.1.8. MAC Frame Format and Addressing

IEEE 802.11 supports three types of frames: management frames, control frames, and data frames. The management frames are used for station association and disassociation with the AP, timing and synchronization, and authentication and Deauthentication. Control frames are used for handshaking and for positive ACKs during the data exchange. Data frames are used for the transmission of data. The MAC header provides information on frame control, duration, addressing, and sequence control. Figure 2.9 illustrates the format of MAC frame [12].

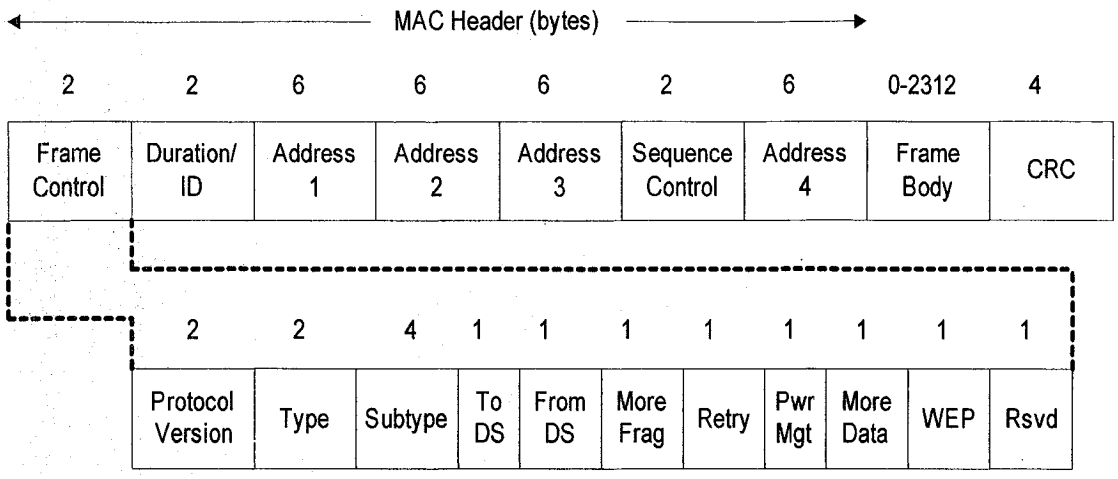

Figure 2.9: MAC Frame Format

2.1.9. Physical (PHY) Layer

The IEEE 802.11 has several physical layers defined to operate with its MAC Layer. Each physical layer is divided into two sublayers that correspond to two protocol functions [10] [12] as shown in Figure 2.10. The Physical Layer Convergence Procedure (PLCP) is the upper sublayer, and it provides a convergence function that maps the MAC PDU into a format suitable for transmission and reception over a given physical medium. The Physical Medium Dependent (PMD) sublayer is concerned with the characteristics and methods for transmitting over the wireless medium. IEEE 802.11 defines three specifications for the Physical Layer: (i) Frequency Hopping Spread Spectrum (FHSS), (ii) Direct Sequence Spread Spectrum (DSSS), and (iii) Infrared (IR).

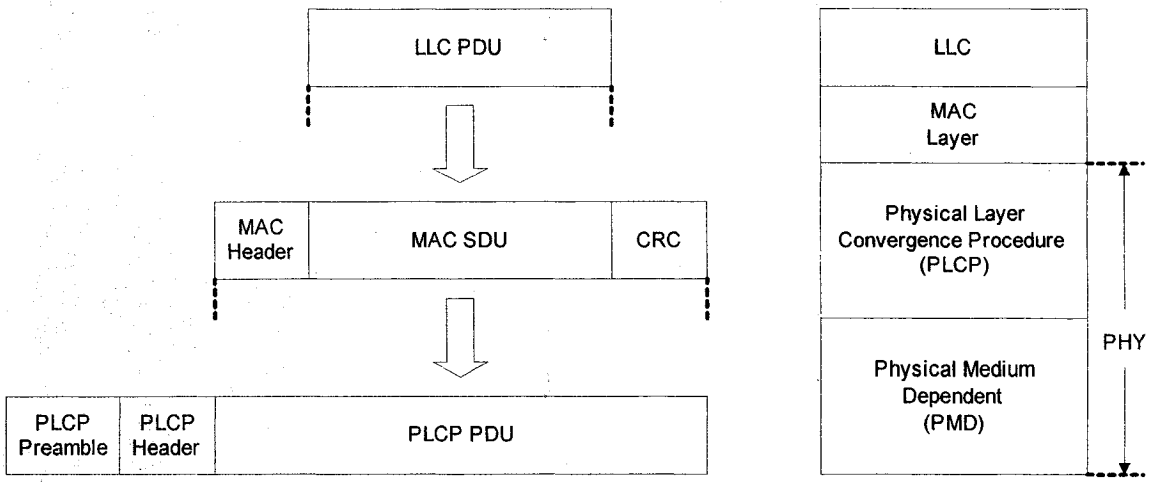

Figure 2.10: IEEE 802.11 Physical Layer Format

2.1.9.1. Frequency Hopping Spread Spectrum (FHSS)

In Frequency Hopping Spread Spectrum (FHSS), the signal is broadcast over a seemingly random series of radio frequencies, hopping from frequency to frequency at fixed intervals as illustrated in Figure 2.11 [10].

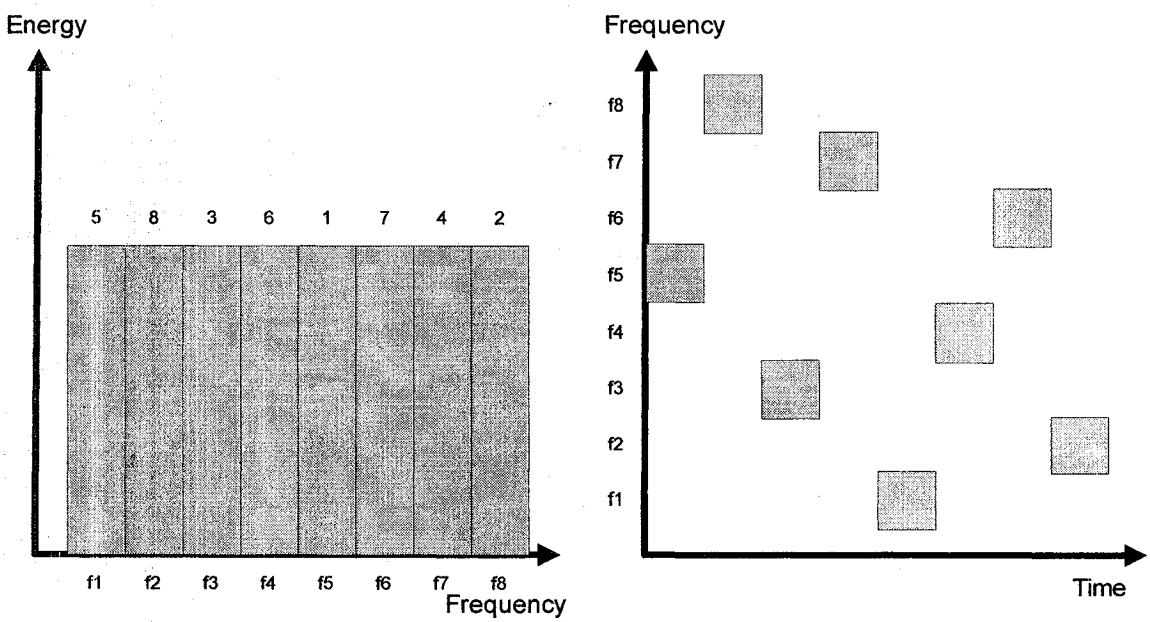

Figure 2.11: Frequency Hopping Spread Spectrum

A hopping code determines the frequencies the radio will transmit and in which order. To properly receive the signal, the receiver must be set to the same hopping code and listen to the incoming signal at the right time and correct frequency. If the radio encounters interference on one frequency, the radio will retransmit the signal on a subsequent hop on another frequency.

The frequency hopping spread spectrum technique reduces interference because an interfering signal from a narrowband system will affect the spread spectrum signal only if both are transmitting at the same frequency at the same time. Therefore the aggregate interference will be very low, resulting in little or no bit errors. In most cases, frequency hopping is the most cost-effective type of WLAN to deploy if needs for network bandwidth are 2 Mbps or less [14].

Frequency hopping may be classified as fast or slow. Fast frequency hopping occurs if there is more than one frequency hop during each transmitted symbol. Thus, fast frequency hopping implies that the hopping rate equals or exceeds the information symbol rate. Slow frequency hopping occurs if one or more symbols are transmitted in the time interval between frequency hops.

2.1.9.2. Direct Sequence Spread Spectrum (DSSS)

A Direct Sequence Spread Spectrum (DSSS) system spreads the baseband data by directly multiplying the basement data pulses with a Pseudo Noise (PN) sequence that is produced by a pseudo noise code generator. In this scheme, each bit to be sent by the

sender is replaced by a sequence of bits called a *chip code*. The ratio between the symbol period and the chip period is called the *processing gain*. A high processing gain increases the signal resistance to interference.

To avoid buffering, the time needed to send one original bit should the same as the time needed to send one chip code. This means that the data rate for sending chip codes should be N times (where N is the number of bits in each chip code) times the data rate of the original bit stream. For example, if the sender generates the original bit stream at 1 Mbps, and the chip code is 7 bits long, the data rate for transferring chip codes should be 1 x 7 = 7 Mbps. Figure 2.12 a typical example. DSSS, having higher potential data rates, would be best for bandwidth-intensive applications.

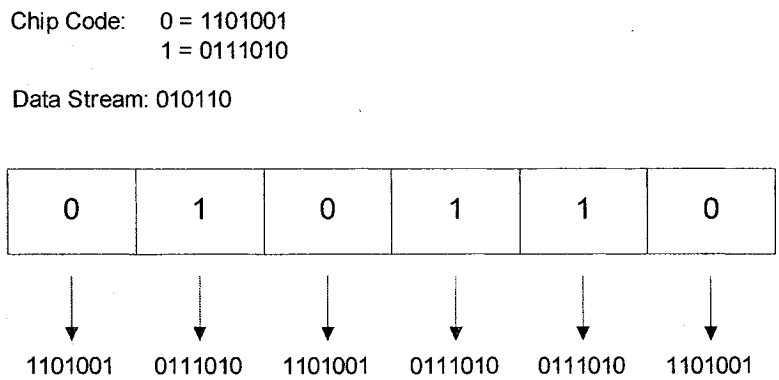

Figure 2.12: Direct Sequence Spread Spectrum

2.1.9.3. Infrared (IR)

The IEEE 802.11 Infrared operates in the near-visible light range of 850 to 890 nanometers. The transmission distance is limited to the range of 10 to 20 meters, and the signal is contained by walls and windows.

2.2. Overview of GPRS

The General Packet Radio Service (GPRS), a standard from the European Telecommunications Standard Institute (ETSI), is designed as an extension to the Global System for Mobile Communication (GSM) network, which provides an efficient way for transporting packet data through wireless channels [18]. GPRS is an evolution of the GSM that uses a time frame structure similar to GSM; an FDMA-TDMA based packet-switched radio technology with 200 KHz channels [23]. However, unlike GSM, it is a data bearer that enables wireless access to data networks like the Internet, enabling users to access E-mail and other Internet applications using mobile phones [53]. The GPRS standard has also been developed for IS-136 [54].

Because GPRS does not require any dedicated end-to-end connection, it only uses network resources and bandwidth when data is actually being transmitted. This means that a given amount of radio bandwidth can be shared efficiently and simultaneously among many users [15]. Another important advantage of GPRS is that it shares physical resources with GSM on a dynamic basis, and makes use of many properties of the Physical Layer of the original GSM system.

The implementation of GPRS has a limited impact on the GSM core network. It simply requires the addition of new packet data switching and gateway nodes, and an upgrades to existing nodes to provide a routing path for packet data between the wireless terminal and a gateway node.

2.2.1. GPRS Network Architecture

A GPRS network architecture [16] is shown in Figure 2.13. The Base Station Controller (BSC) is equipped with the Packet Control Unit (PCU), which supports all GPRS protocols for communication over the air interface. PCU supports cell change, radio resource configuration, and channel assignment. Besides, the Home Location Register (HLR) is enhanced with GPRS subscriber data and routing information [17].

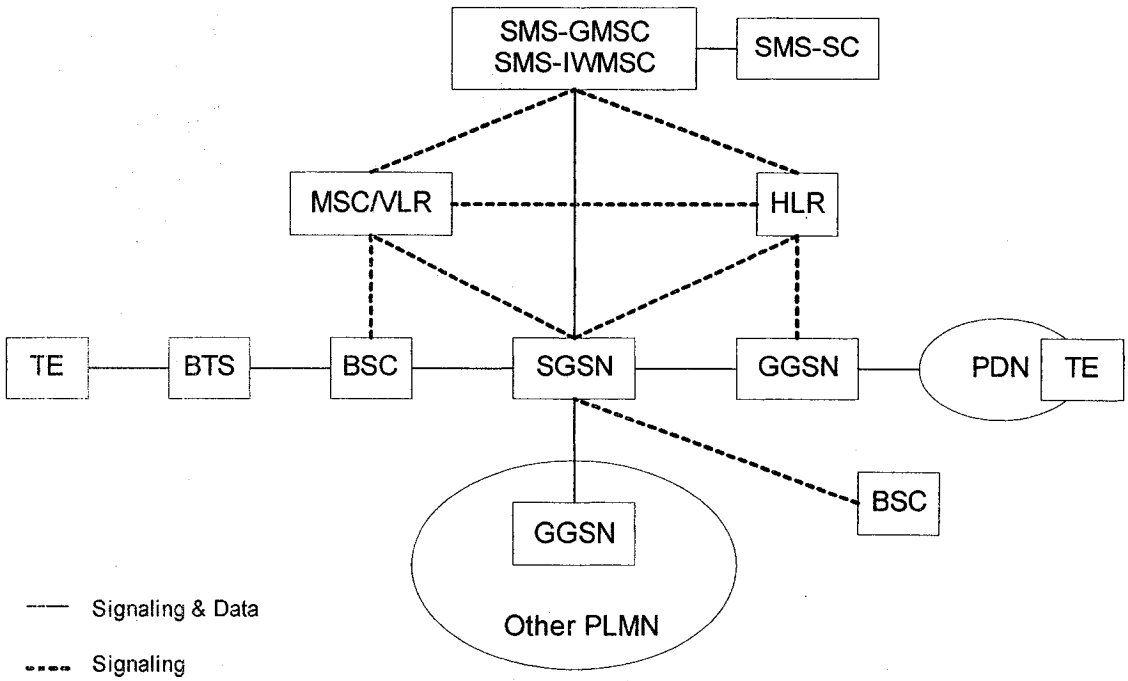

Figure 2.13: GPRS Network Architecture

Independent packet routing and transfer within the Public Land Mobile Network (PLMN) is supported by a new logical network node called the GPRS Support Node (GSN). There are two GSNs in the GPRS system - Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). The GGSN acts as a logical interface to external Packet Data Networks (PDN); e.g. Internet. The SGSN is responsible for the delivery of packets

to the Mobile Stations (MSs) within its service area. Within the GPRS networks, PDUs are encapsulated at the original GSN and decapsulated at the desination GSN. In between the GSNs, IP is used as the backbone to transfer PDUs. This whole process is referred to as tunneling [15] [18].

2.2.2. Packet Data Routing

Figure 2.14 show routing of packet data in a mobile originated transmission. The SGSN of the source mobile encapsulates the packets transmitted by the MS and routes them to the appropriate GGSN. Based on the examination of the destination address, packets are then routed to the destination GGSN through the PDN [18].

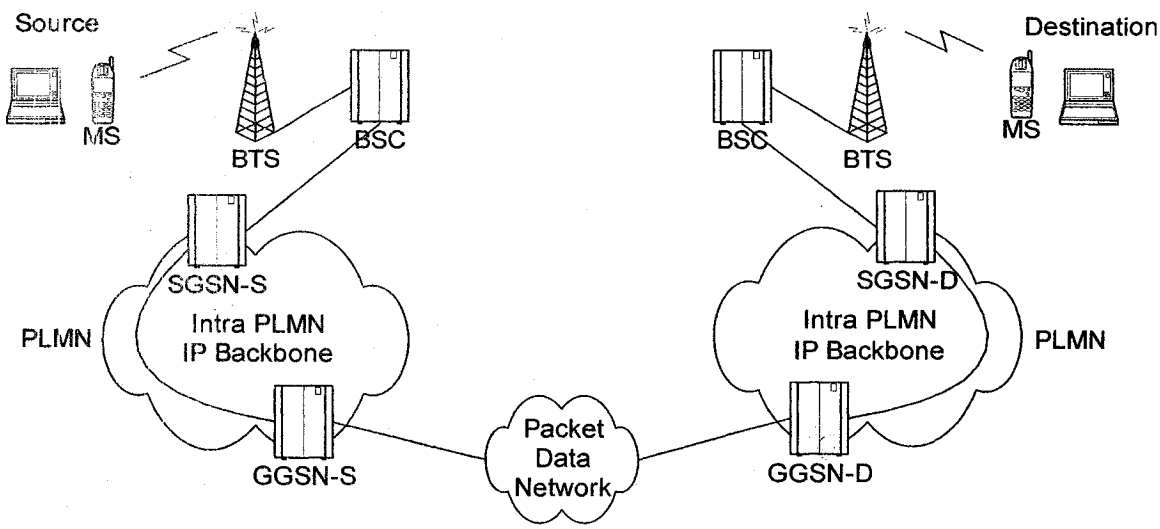

Figure 2.14: Routing of Packet Data in GPRS Networks

The destination GGSN checks the routing context associated with the destination address and determines the destination SGSN and relevant tunneling information. Each packet is

then decapsulated and forwarded to the destination SGSN, which delivers it to the destination mobile.

2.2.3. GPRS Protocol Architecture

The protocol stack of GPRS network is shown in Figure 2.15 in a simplified form according to the OSI Reference Model. Above the Network Layer, widespread standardize protocols may be used [18].

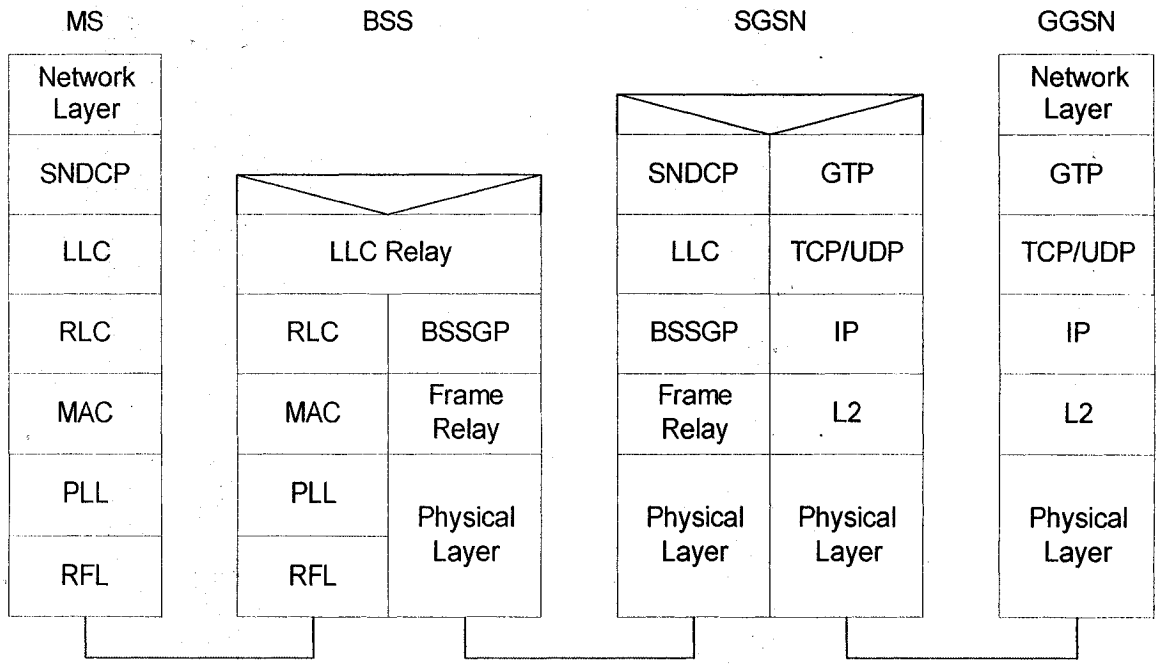

Figure 2.15: GPRS Protocol Stack

Between two GSNs, GPRS Tunnel Protocol (GTP) tunnels the PDU through the GPRS backbone network by adding routing information. Below the GTP, the Transmission Control Protocol / User Datagram Protocol (TCP/UDP) and the Internet Protocol (IP) are used as the GPRS network-layer backbone protocols. Ethernet, Integrated Service Digital

Network (ISDN), or Asynchronous Transfer Mode (ATM) based protocols may be used below IP.

Between SGSN and MS, the Sub Network Dependent Convergence Protocol (SNDCP) maps network level protocol characteristics onto the underlying LLC and provide functionalities like multiplexing of network-layer messages onto a single virtual logical connection, encryption, segmentation, and compression.

Between the MS and BSS, the Data Link Layer has been separated into two distinct sublayers: Logical Link Control (LLC) and Radio Link Control / Medium Access Control (RLC/MAC) sublayers. The LLC layer provides a logical link between the MS and the SGSN.

The RLC/MAC layer provides services for information transfer over the Physical layer of the GPRS radio interface. It defines the procedure that enables multiple MSs to share a common transmission medium which may consist of several physical channels. The RLC layer is responsible for the transmission of data blocks across the air interface. The MAC layer, derived from a slotted ALOHA-protocol, is responsible for access signaling procedures for the radio channel governing the attempts to access the channel by the MSs, and the control of that access by the network side.

The Physical layer is split up into a Physical Link Sublayer (PLL) and a Physical RF Sublayer (RFL). The PLL is responsible for: Forward Error Correction (FEC) coding,

rectangular interleaving, and detecting physical link congestion. The RFL performs modulation and demodulation of the physical waveforms.

In the network, the LLC is split between the BSS and SGSN. The BSS functionality is called LLC Relay. Between the BSS and SGSN, the BSS GPRS Protocol (BSSGP) conveys routing and Quality of Service (QoS)-related information, and operates over Frame Relay [18].

2.2.4. Frame Size, Bit Rate, and Coding Schemes

The data rate offered by GPRS system depends on how many time slots are selected by the MH and also the radio channel condition [45]. A basic data unit transferred on GPRS Packet Data Channel (PDCH) is an RLC block. It is transmitted during one block period, which is a sequence of four time slots on a PDCH [19]. The RLC data block consists of a RLC header, RLC data field, and spare bits. Each RLC data block may be encoded using any of the available channel coding schemes CS1, CS2, CS3, or CS4, and thus the user data size of an RLC block depends on the channel coding chosen. The size of the RLC data block and the data rate for each of the channel coding schemes [15] [19] is shown in Table 1.

Each dedicated channel is divided into eight time slots, with each time slot supporting a maximum data transmission speed of 21.4 Kbps. The theoretical maximum data rate is 171.2 Kbps per channel. However, the user will not experience anything close to this because the data rate assumes no error correction and the use of all eight time slots.

Normal data transmission utilizes error correction, which limits the data rate per time slot to 13.4 Kbps [20].

Coding Scheme	*Code Rate*	*RLC Data Block Size (bits)*	*Data Rate (Kbps)*
CS-1	1/2	181	9.05
CS-2	~ 2/3	268	13.4
CS-3	~ 3/4	312	15.6
CS-4	1	428	21.4

Table 2.1: GPRS Coding Schemes

The selection of coding schemes is transparent to the user and determines the level of error correction the network uses to send the data. The better the link is between the user and the network, the less error correction is needed [21]. Less error correction means higher throughput. (CS-1 has the highest level of error correction.) For example, CS-2 provides a rate of 13.4 Kbps per slot. The number of slots is simply multiplied by 13.4 Kbps to determine the aggregated data rate.

2.2.5. Packet Size and Overhead

In our discussion we are considering the voice calls to be processed by the GPRS core packet network, and thus using the Internet Protocol (IP). IP Protocol provides only means to transmit data packets between source and destination. Because the transmission delay may vary from packet to packet, transmitted speech (voice) frame must be provided with timing information to ensure correct reconstruction of the data stream in the

receiver. Therefore, in voice over IP, it is commonly assumed that speech data is further encapsulated by Real Time Protocol (RTP) and User Datagram Protocol (UDP) protocols. Figure 2.16 shows the frame (voice) size and overhead assuming CS-1 and one voice frame per RTP packet [19].

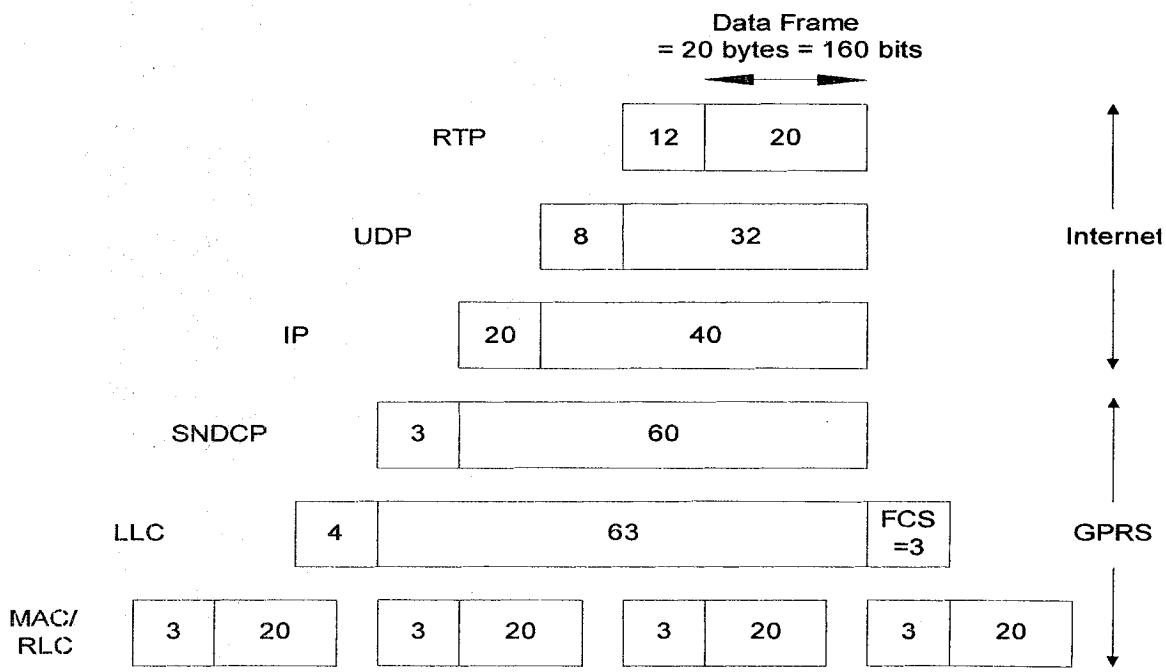

Figure 2.16: GPRS Frame (voice) Size and Overhead (in bytes)

Following is the length of header being used in IP voice over GPRS [22]:

User Datagram Protocol (UDP)	= 8 bytes
Real Time Protocol (RTP)	= 12 bytes
Internet Protocol (IP)	= 20 bytes [for IPv4]
Sub-Network dependent Convergence Protocol (SNDCP)	= 3 bytes
Logical Link Control (LLC)	= 7 bytes

Finally, when LLC packet is divided into approximate number of RLC blocks, each block has a 3-4 byte MAC+RLC header. Thus, total header size following from different protocols required for IP voice over GPRS is well over 50 bytes. Typically, the size of an encoded speech frame varies from 10-35 bytes. This means that only less than 50% of the transmitted data comes from the speech source. In the above example, the speech frame has been considered as of 20 bytes for CS-1 coding scheme. However, the frame size is 30 bytes for CS-2 [22].

There are two basic methods to improve relatively poor efficiency caused by excess of different header information. Firstly, if more than one speech frame is encapsulated into a packet, the relative header size compared to payload decreases. Secondly, effective algorithm for UDP/RTP/IP header compression has been developed and it has been shown that the 40-byte header can quite easily be compressed to 2-4 bytes. However, both these approaches have drawbacks. Encapsulating several speech frames into a single packet increases the end-to-end delay making conversation difficult. Additional delays should be avoided because there are also several other contributors to the total end-to-end delay in the IP voice over GPRS scheme. The most effective header compression schemes may not be robust to transmission errors. However, header compression is perhaps still a better way to decrease the overhead. Especially because recently proposed schemes provide fairly reliable and efficient compression with the expense of some added complexity [22].

2.2.6. GPRS Mobility Management

The operation of the GPRS is partly independent of the GSM network. However, some procedures share the network elements with current GSM functions to increase efficiency and to make optimum use of free GSM resources (such as unallocated time slots). An MS has three states in the GPRS system: idle, standby, and active. The three-state model represents the nature of packet radio relative to the GSM two-state model (idle or active) [23]. We discuss two types of call scenario.

2.2.6.1. MS Terminated Call

Data is transmitted from GPRS network to MS only when the MS is in the active state. In the active state, the SGSN knows the cell location of the MS. However, in the standby state, the location of the MS is only known as to which routing area it is in. (The routing area can consist of one or more cells within a GSM location area.) When the SGSN sends a packet to a MS that is in the standby state, the MS must be paged. Because the SGSN knows the routing area in which the MS is located, a packet paging message is sent to that routing area. After receiving the packet paging message, the MS gives its cell location to the SGSN to establish the active state. Packet transmission to an active MS is initiated by packet paging to notify the MS of an incoming data packet. The data transmission proceeds immediately after packet paging through the channel indicated by the paging message. The purpose of the packet paging message is to simplify the process of receiving packets. The MS has to listen to only the packet paging messages, instead of all the data packets in the downlink channels, reducing battery use significantly.

2.2.6.2. MS Originated Call

When an MS has a packet to be transmitted, access to the uplink channel is needed. The uplink channel is shared by a number of MSs, and its use is allocated by a BSS. The MS requests use of the channel in a packet random access message. The transmission of the packet random access message follows Slotted Aloha procedures. The BSS allocates an unused channel to the MS and sends a packet access grant message in reply to the packet random access message. The description of the channel (one or multiple time slots) is included in the packet access grant message. The data is transmitted on the reserved channels.

2.2.6.3. Routing Update (RU)

The main reasons for the standby state are to reduce the load in the GPRS network caused by cell-based routing update messages and to conserve the MS battery. When a MS is in the standby state, there is no need to inform the SGSN of every cell change—only of every routing area change. The operator can define the size of the routing area and, in this way, adjust the number of routing update messages.

In the idle state, the MS does not have a logical GPRS context activated or any Packet-Switched Public Data Network (PSPDN) addresses allocated. In this state, the MS can receive only those multicast messages that can be received by any GPRS MS. Because the GPRS network infrastructure does not know the location of the MS, it is not possible to send messages to the MS from external data networks [23].

2.2.6.3.1. RU for Inter Cell Movement: A cell-based routing update procedure is invoked when an active MS enters a new cell. In this case, the MS sends a short message containing information about its move (the message contains the identity of the MS and its new location) through GPRS channels to its current SGSN. This procedure is used only when the MS is in the active state.

2.2.6.3.2. RU for Inter Routing Area Movement: When an MS in an active or a standby state moves from one routing area to another in the service area of one SGSN, it must again perform a routing update. The routing area information in the SGSN is updated and the success of the procedure is indicated in the response message.

2.2.6.3.3. RU for Inter SGSN Movement: The inter-SGSN routing update is the most complicated of the three routing updates. In this case, the MS changes from one SGSN area to another and it must establish a new connection to a new SGSN. This means creating a new logical link context between the MS and the new SGSN, as well as informing the GGSN about the new location of the MS.

2.3. Overview of Mobile IP

The Internet infrastructure is built on top of a collection of protocols, called the TCP/IP protocol suite. Transmission Control Protocol (TCP) and Internet Protocol (IP) are the core protocols in this suite. IP routes packets to their destinations according to IP addresses. These addresses are associated with a fixed network location. When the packet's destination is a mobile node, this means that each new point of attachment made by the node is associated with a new network number and, hence, a new IP address, making transparent mobility impossible [24]. Mobile IP, a standard proposed by a working group within the Internet Engineering Task Force (IETF) in RFC 2002 [25], was designed to solve this problem by allowing the mobile node to use two IP addresses.

2.3.1. TCP/IP Network Architecture

Unlike the seven-layer Open System Interconnection (OSI) architecture, TCP/IP network architecture is a four-layer system in which each layer is responsible for a specific task [12]. The four layers, from top to bottom, are Application Layer, Transport Layer, Internet Layer, and Network Interface Layer, as shown in Figure 17(a).

The Application Layer handles the details of the particular application (e.g., FTP, TELNET, HTTP etc.). The Transport Layer provides a flow of data between two Internet nodes. There are two widely used transport layer protocols on the Internet: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). TCP provides a reliable flow of data between two nodes by maintaining a connection-oriented environment. On the other hand, UDP provides an unreliable and connectionless datagram service. The

Internet Layer handles the movement of packets around the network by implementing efficient routing algorithms. The Network Interface Layer provides interfaces to the network hardware devices, and is concerned with the network-specific aspects of the transfer of packets. Examples include IEEE 802.2 LANs, X.25, Frame Relay etc. The TCP/IP model does not require strict layering [12], as shown in Figure 17(b).

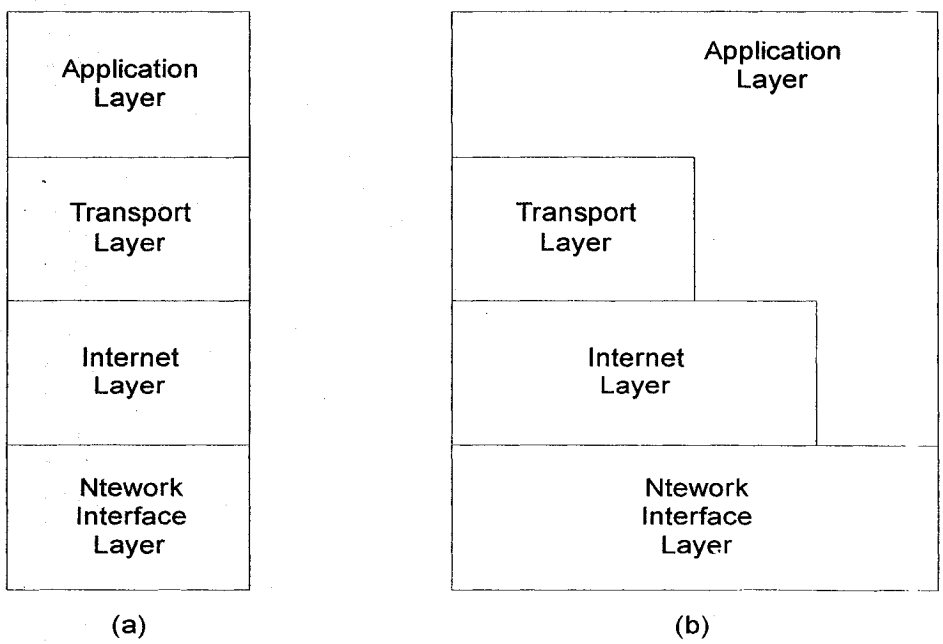

Figure 2.17: TCP/IP network Architecture

2.3.1.1. Brief Overview of Internet Protocol: IPv4 and IPv6

The Internet Protocol (IP) is the heart of the TCP/IP protocol suite. IP corresponds to the Network Layer in the OSI Reference Model, and provides a connectionless and best effort delivery service to the Transport Layer. IP is designed for use in interconnected systems of packet-switched computer communication networks [26]. The internet protocol provides the transmission of data blocks called IP Packets (or sometimes called

IP Datagrams) from sources to destinations, where sources and destinations are hosts identified by fixed length addresses. IP packets are sent to the Network Interface Layer for delivery across the physical network. At the receiver, packets passed up by the Network Interface Layer are demultiplexed to the Internet Layer as shown in Figure 2.18.

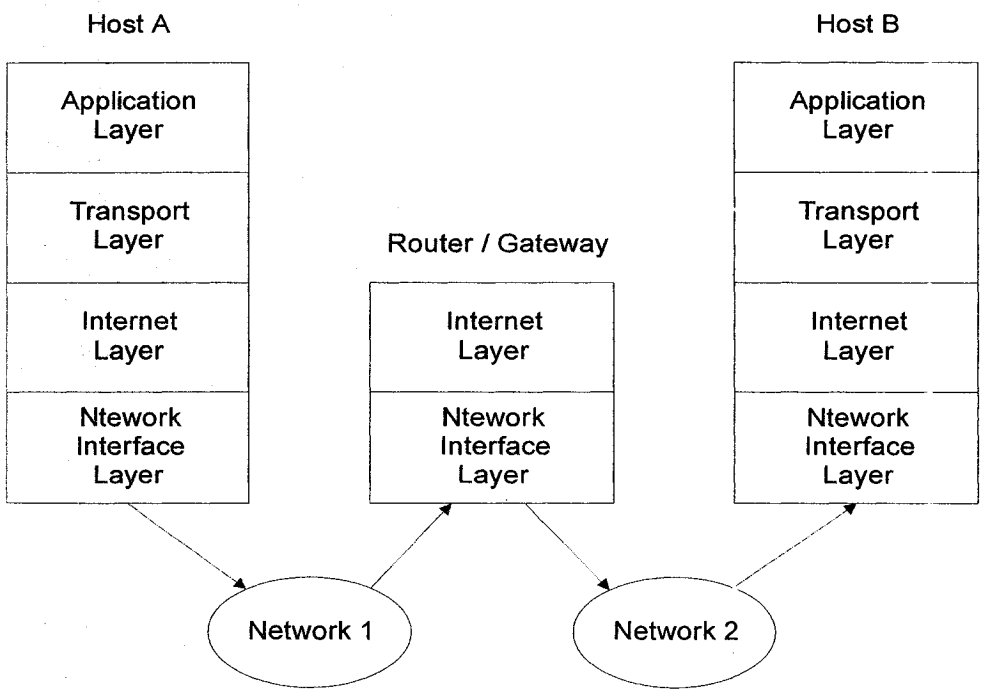

Figure 2.18: The Internet and Network Interface Layer

The IP implements two basic functions: addressing and fragmentation. The internet modules use the addresses carried in the internet header to transmit internet datagrams toward their destinations. Each host in the Internet is identified by a globally unique IP address. The selection of a path for transmission is called routing. The internet modules use fields in the internet header to fragment and reassemble internet datagrams when necessary.

The initial version of IP, known as IP version 4 (or IPv4), has played a central role in internetworking environment for many years. However, it has become a victim of its own success; i.e. explosive growth [12]. In the 1990s, the World Wide Web and personal computers shifted the user of the Internet from researchers and scientists to the general public. This change has created heavy demands for new IP addresses, and the 32 bits of the current IP addresses will be exhausted sooner or later.

In the early 1990s, the IETF began to work on the successor of IPv4 that would solve the address exhaustion and other scalability problem. The new IP version, known as IP version 6 (or IPv6) was recommended in late 1994. IPv6 was designed to interoperate with IPv4 since it would likely take many years to complete the transition from version 4 to version 6. IPv6 should also change the IPv4 functions that do not work well and support new emerging applications such as real-time video conferencing, etc.

2.3.2. Motivation for Mobile IP

IP routes packets from a source endpoint to a destination by allowing routers to forward packets from incoming network interfaces to outbound interfaces according to routing tables. The routing tables typically maintain the next-hop information for each destination IP address, according to the number of networks to which that IP address is connected. The network number is derived from the IP address by masking off some of the low-order bits. Thus, the IP address typically carries with it information that specifies the IP node's point of attachment.

To maintain existing transport-layer connections as the mobile node moves from place to place, it must keep its IP address the same. In TCP connections are indexed by a quadruplet that contains the IP addresses and port numbers of both connection endpoints. Changing any of these four numbers will cause the connection to be disrupted and lost. On the other hand, correct delivery of packets to the current point of attachment of the Mobile Hosts (MH) depends on the network number contained within the MH's IP address, which changes at new points of attachment. To change the routing requires a new IP address associated with the new point of attachment.

2.3.3. Mobile IP Operation

Mobile IP has been designed to solve the problem of mobility in IP by allowing the MH to use two IP addresses. In Mobile IP, the Home Address is static and is used to identify TCP connections. The Care-of Address (CoA) changes at each new point of attachment and can be thought of as the MH's topologically significant address; it indicates the network number and thus identifies the MH's point of attachment with respect to the network topology. The Home Address makes it appear that the MH is continually able to receive data on its home network, where Mobile IP requires the existence of a network node known as the Home Agent (HA). Whenever the MH is not attached to its home network (and is therefore attached to what is termed a Foreign Network), the HA gets all the packets destined for the MH and arranges to deliver them to the MH's current point of attachment; i.e Foreign Agent (FA). The HA maintains the mobility binding in a Mobility Binding Table, while the FA maintains a Visitor List which contains information about the MH currently visiting that network [27] as shown in Figure 2.19.

Home Address	Care-of Address	Life Time (in sec)
129.191.170.4	126.170.22.78	200
129.191.170.2	117.121.55.77	150

(a)

Home Address	HA Address	Media Address	Life Time (in sec)
129.191.42.13	129.191.42.7	00-60-08-95-66-E1	200
129.191.31.18	129.191.31.1	00-60-08-68-A2-56	150

(b)

Figure 19: (a) Mobility Binding Table, (b) Visitor List

Whenever the MH moves, it registers its new CoA with its HA. To get a packet to a MH from its home network, the HA delivers the packet from the home network to the CoA. The further delivery requires that the packet be modified so that the CoA appears as the destination IP address. This modification can be understood as a packet transformation or, more specifically, a redirection. When the packet arrives at the CoA, the reverse transformation is applied so that the packet once again appears to have the MH's home address as the destination IP address. When the packet arrives at the MH, addressed to the home address, it will be processed properly by TCP.

In Mobile IP, the HA redirects packets from the home network to the CoA by constructing a new IP header that contains the MH's CoA as the destination IP address. This new header then shields or encapsulates the original packet, causing the MH's home address to have no effect on the encapsulated packet's routing until it arrives at the CoA. Such encapsulation is also called tunneling [25], which is shown in Figure 2.20.

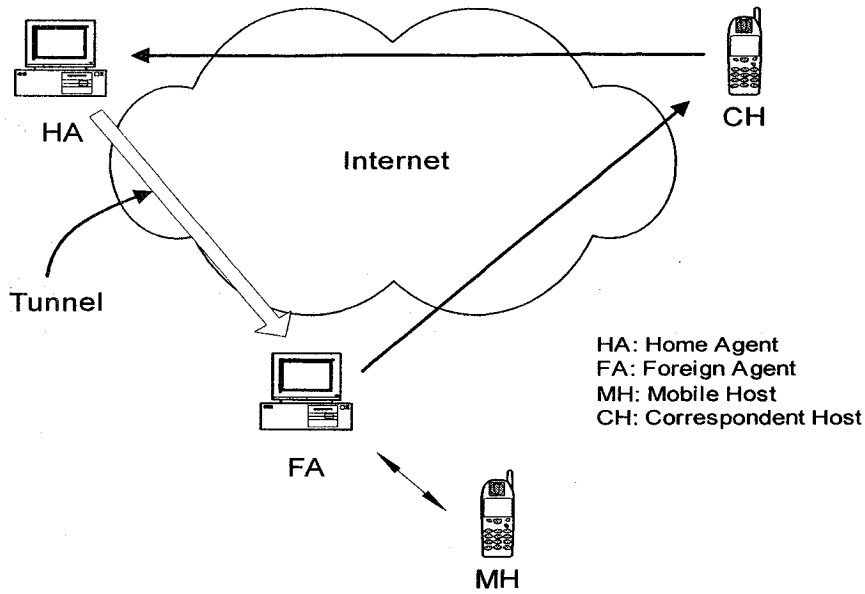

Figure 2.20: Operation of Mobile IP

Mobile IP can be thought of as the cooperation of three major subsystems [28]: Agent Discovery, Registration, and Tunneling.

2.3.3.1. Mobile Agent Discovery

The Mobile IP discovery process has been built on top of an existing standard protocol, Internet Control Message Protocol (ICMP) Router Advertisement, specified in RFC 1256 [29]. Mobile IP discovery does not modify the original fields of existing router advertisements but simply extends them to associate mobility functions. The format of the main messages related to agent discovery is shown in Figure 2.21 through 2.24.

2.3.3.1.1. Agent Advertisement: When the router advertisements are extended to also contain the needed CoA, they are known as Agent Advertisements. HAs and FAs typically broadcast agent advertisements at regular intervals (e.g, once a second or once

every few seconds) in a random fashion [30]. If sent periodically, the nominal interval at which Agent Advertisements are sent should be 1/3 of the advertisement Lifetime given in the ICMP header. This allows a MH to miss three successive advertisements before deleting the agent from its list of valid agents [25].

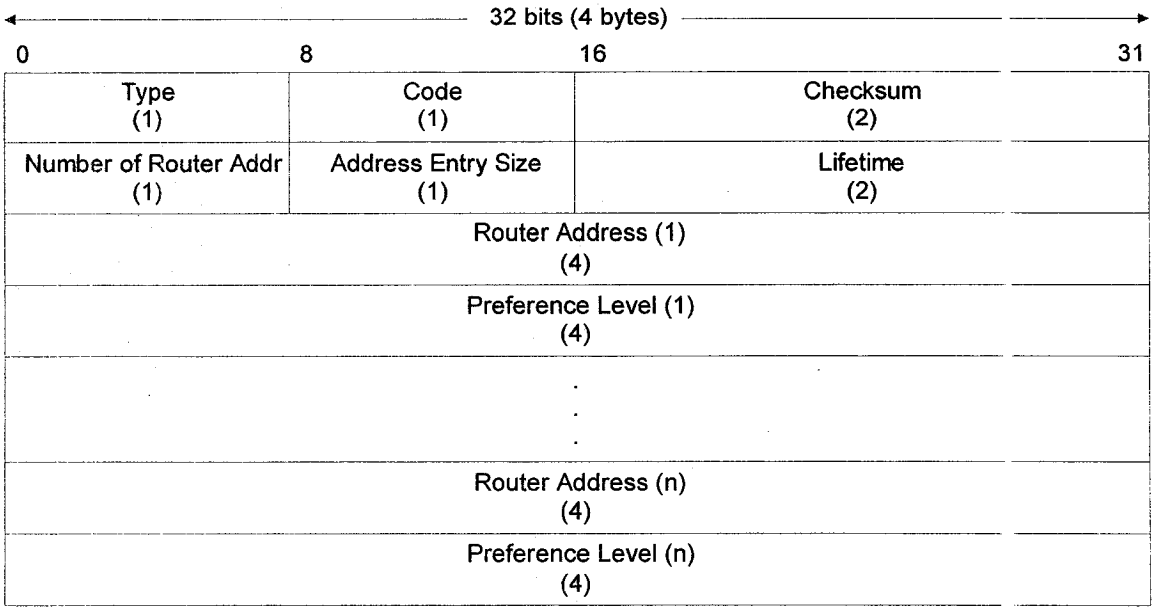

Figure 2.21: Router Advertisement (from RFC 1256)

*2.3.3.1.2. **Mobility Agent Advertisement Extension:*** The Mobility Agent Advertisement Extension follows the ICMP Router Advertisement fields. It is used to indicate that an ICMP Router Advertisement message is also an Agent Advertisement being sent by a mobility agent [25].

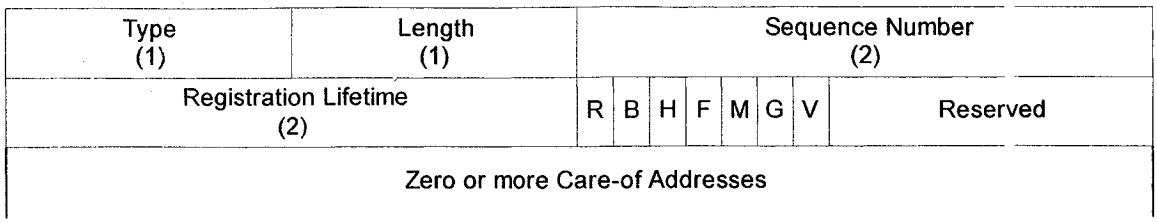

Figure 2.22: Mobility Agent Advertisement Extension

2.3.3.1.3. Prefix-Lengths Extension: The Prefix-Lengths Extension may follow the Mobility Agent Advertisement Extension. It is used to indicate the number of bits of network prefix that applies to each Router Address listed in the ICMP Router Advertisement portion of the Agent Advertisement.

Figure 2.23: Prefix Length Extension

2.3.3.1.4. Agent Advertisement Solicitation: If a MH needs to get a CoA and does not wish to wait for the periodic advertisement, the MH can broadcast or multicast a solicitation, known as Agent Advertisement Solicitation message, that will be answered by any FA or HA that receives it.

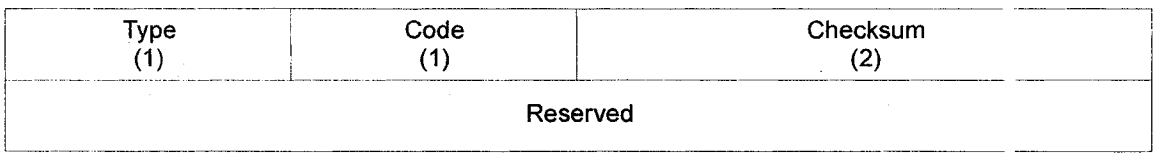

Figure 2.24: Router Solicitation (from RFC 1256)

2.3.3.2. Registration

Mobile IP defines two different registration procedures, one via a FA that relays the registration to the MH's HA, and one directly with the MH's HA. Figure 2.25 illustrates the registration process via FA. The Mobile IP registration messages use the UDP [31].

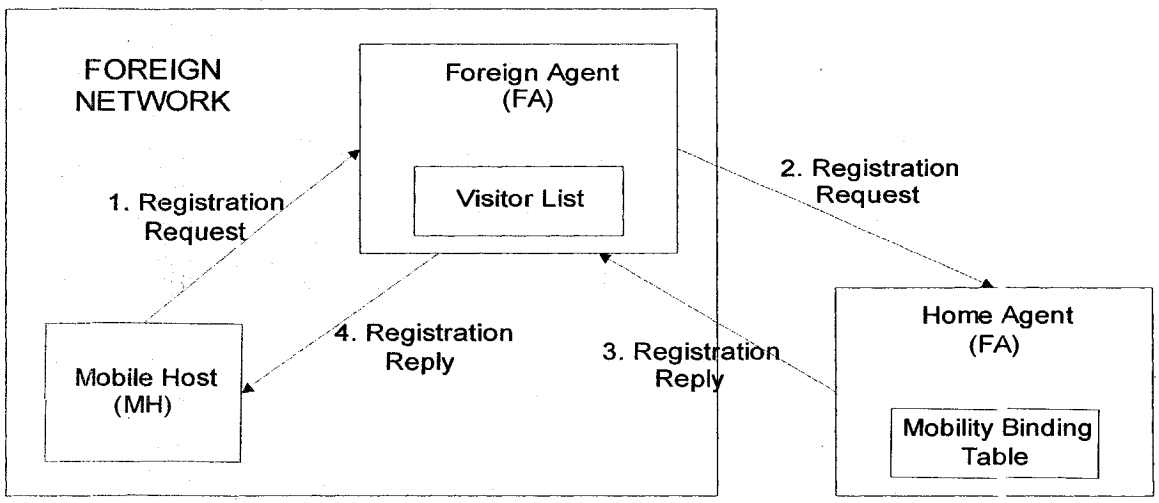

Figure 2.25: Mobile IP Registration

*2.3.3.2.1. **Registration Request:*** A MH registers with its HA using a Registration Request message so that its HA can create or modify (e.g., with a new lifetime) the Mobility Binding Table for that MH. Figure 2.26 exhibits the format of Registration Request message.

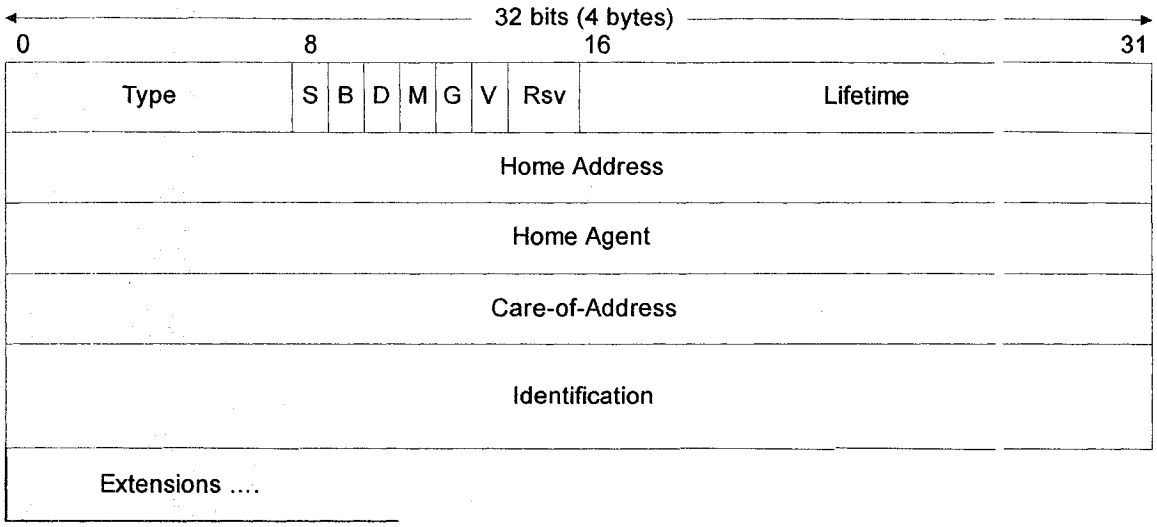

Figure 2.26: Registration Request (from RFC 2002)

2.3.3.2.2. Registration Reply: A mobility agent returns a Registration Reply message to a MH which has sent a Registration Request message. If the MH is requesting service from a FA, that FA will receive the Reply from the HA and subsequently relay it to the MH. Figure 2.27 exhibits the format of Registration Request message.

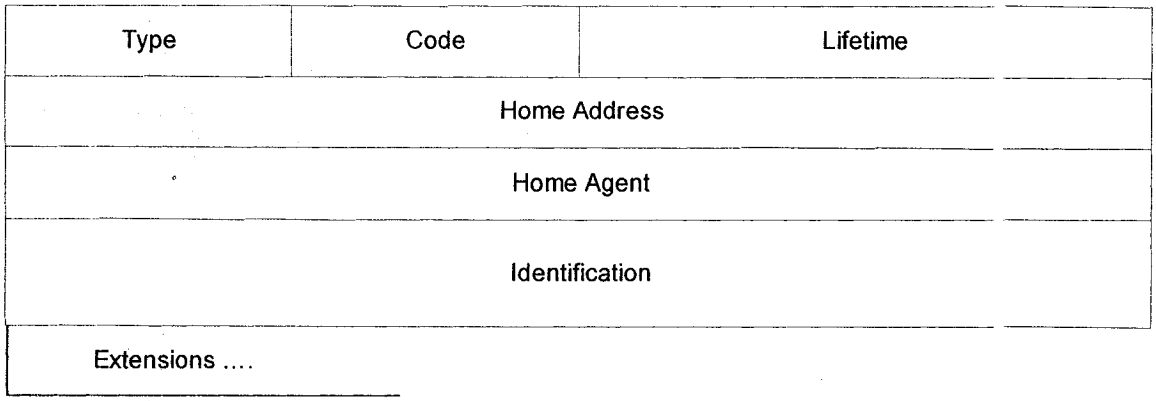

Figure 2.27: Registration Reply (from RFC 2002)

*2.3.3.2.3. **Authentication:*** The HA must be certain that registration was originated by the MH and not by some other malicious node pretending to be the MH. A malicious node could cause the HA to alter its routing table with erroneous CoA information, and the MH would be unreachable to all incoming communications from the Internet. The need to authenticate registration information has played a major role in determining the acceptable design parameters for Mobile IP. Each MH and HA must share a security association and be able to use Message Digest 5 with 128-bit keys to create unforgeable digital signatures for registration requests [27] [32]. The signature is computed by performing MD5's one-way hash algorithm over all the data within the registration message header and the extensions that precede the signature. The fixed portion of the Registration Reply is followed by one or more of the Extension. The Mobile-Home Authentication Extension must be included in all Registration Replies returned by the HA [31] as shown in Figure 2.28.

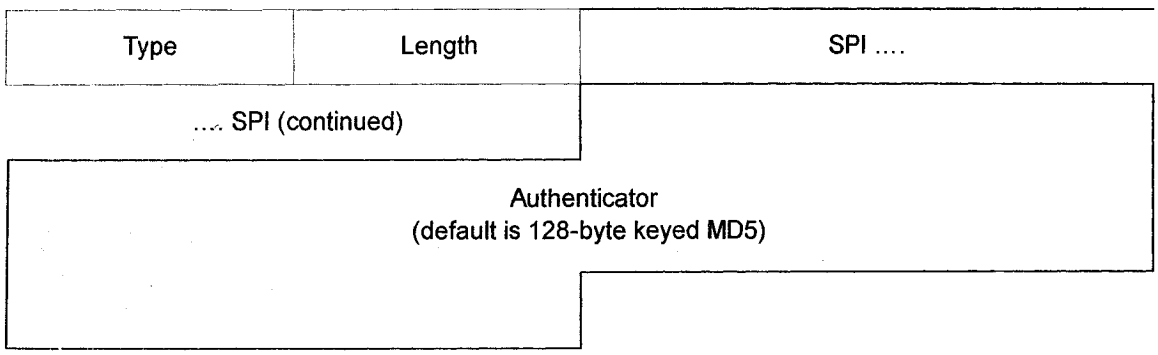

Figure 2.28: Authentication Extension (from RFC 2002)

2.3.3.3. Tunneling

When a Correspondent Host (CH) wants to communicate with the MH, it sends an IP packet addressed to the permanent IP address of the MH. The HA intercepts this packet and consults the Mobility Binding Table to find out if the MH is currently visiting any other network. The HA finds out the mobile node's CoA and constructs a new IP header that contains the MH's CoA as the destination IP address. The original IP packet is put into the payload of this IP packet. It then sends the packet. This process of encapsulating one IP packet into the payload of another is known as IP-within-IP encapsulation [33], or tunneling, as shown in Figure 2.29.

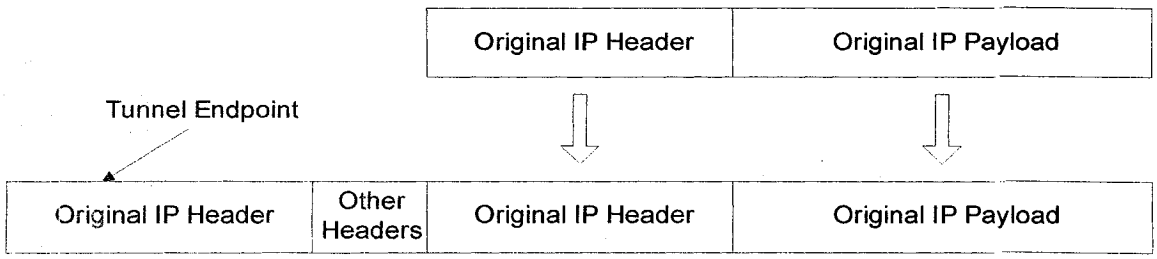

Figure 2.29: IP-in-IP Encapsulation or Tunneling

When the encapsulated packet reaches the MH's current network, the FA decapsulates the packet and finds out the MH's home address. It then consults the visitor list to see if it has an entry for that MH. If there is an entry for the MH on the visitor list, the FA retrieves the corresponding media address and relays it to the MH.

When the MH wants to send a message to a CH, it forwards the packet to the FA, which in turn relays the packet to the correspondent node using normal IP routing. The FA

continues serving the MH until the granted lifetime expires. If the MH wants to continue the service, it has to reissue the Registration Request.

2.3.4. Triangular Routing and Route Optimization

Packets sent to a MH are routed indirectly through the MH's HA and are then tunneled to the MH, whereas packets sent from a MH are routed directly to the CH. Roundtrip communications thus travel along three distinct paths, and this routing anomaly is known as Triangle Routing, as shown in Figure 2.20. This method may be inefficient in many cases. If we consider the case when the CH and the MH are in the same network, but not in the home network of the MH. In this case the messages will experience unnecessary delay since they have to be first routed to the HA that resides in the home network. One way to improve this is through Route Optimization.

Route Optimization is an extension proposed to the basic Mobile IP protocol [34]. Here messages from the CH are routed directly to the MH's CoA without having to go through the HA. Route Optimization provides the following four main operations: (i) updating binding caches, (ii) managing smooth handoffs between foreign agents, (iii) acquiring registration keys for smooth handoffs, and (iv) using special tunnels [31].

2.3.5. Minimal Encapsulation

Encapsulation in Mobile IP is carried out by putting the original datagram (= IP header + payload) inside another IP envelope. The fields in the outer IP header add too much overhead to the final datagram; several fields are duplicated from the inner IP header. To

prevent this waste of space a Minimal Encapsulation Scheme [35] has been defined where instead of inserting a new header, the original header is modified to reflect the CoA and a minimal forwarding header is inserted to store the original source and destination addresses. Thus the CoA of the MH becomes the destination address of the IP packet and the HA's address becomes the source address. The minimal forwarding header stores the original source and destination addresses. When the FA tries to decapsulate, it simply restores the fields in the forwarding header to the IP header and removes the forwarding header.

2.3.6. Changes in IPv6

IPv6 includes many features for streamlining mobility support that are missing in IPv4, including Stateless Address Autoconfiguration and Neighbor Discovery. IPv6 also attempts to drastically simplify the process of renumbering, which could be critical to the future routability of the Internet.

Mobility Support in IPv6, as proposed by the Mobile IP working group, follows the design for Mobile IPv4 [28]. It retains the ideas of a home network, HA, and the use of encapsulation to deliver packets from the home network to the MH's current point of attachment. While discovery of a CoA is still required, a MH can configure its CoA by using Stateless Address Autoconfiguration and Neighbor Discovery. Thus, FAs are not required to support mobility in IPv6 [24].

2.3.6.1. Route Optimization:

IPv6 mobility borrows heavily from the route optimization ideas specified for IPv4 [34], particularly the idea of delivering binding updates directly to CH. When it knows the MH's current CoA, a CH can deliver packets directly to the MH's home address without any assistance from the HA. Route optimization is likely to dramatically improve performance for IPv6 MHs.

2.3.6.2. Security:

One of the biggest differences between IPv6 and IPv4 is that all IPv6 hosts are expected to implement strong authentication and encryption features to improve Internet security [31]. This affords a major simplification for IPv6 mobility support, since all authentication procedures can be assumed to exist when needed and do not have to be specified in the Mobile IPv6 protocol.

2.3.6.3. Source Routing:

In contrast to the way in which route optimization is specified in IPv4, in IPv6 CHs do not tunnel packets to MHs. Instead, they use IPv6 routing headers, which implement a variation of IPv4's source routing option. Other features supported by IPv6 mobility include [31]:

- coexistence with Internet ingress filtering;
- smooth handoffs;
- renumbering of home networks; and
- automatic home agent discovery.

Chapter 3

3. Handoff Between WLAN and GPRS

3.1. Handoff and Signal Strength

When a Mobile Host (MH) moves away from a point of attachment, the signal level degrades and there is a need to switch communications to another point of attachment. Handoff is the mechanism by which an ongoing connection between a Mobile Host (MH) and a Correspondent Host (CH) is transferred from one point of attachment to another. In cellular telephony, such point of attachments are referred to as Base Stations (BSs) and in Wireless Local Area Networks (WLANs), they are called Access Points (APs). In either case, such point of attachment serves a coverage area called a Cell.

Handoff, in case of cellular telephony, involves the transfer of a call from one BS to another. In case of WLANs it involves transferring the connection from one AP to another. In a heterogeneous network of cellular and WLAN technology, it will involve the transfer of connection from one BS to another, from one AP to another, between a BS and an AP, or vice versa.

Handoffs must be performed successfully and as infrequently as possible, and be imperceptible to all users. Handoff algorithms are conventionally based on measurement of the Received Signal Strength (RSS). Once a particular signal level is specified as the minimum usable signal for acceptable voice quality at the base station receiver (normally

taken as between -90 dBm and -100 dBm), a slightly stronger signal level is used as a threshold at which a handoff is made [36].

This margin, given by $\Delta = P_{RSS(handoff)} - P_{RSS(minimum\ usable)}$, cannot be too large or too small. If Δ is too large, unnecessary handoffs which burden the network may occur. On the other hand, if Δ is too small, there may be insufficient time to complete a handoff before a call is lost due to weak signal condition. Therefore, Δ is chosen carefully to meet these conflicting requirements. Figure 3.1 illustrates two handoff situations as the MH crosses the cell boundary.

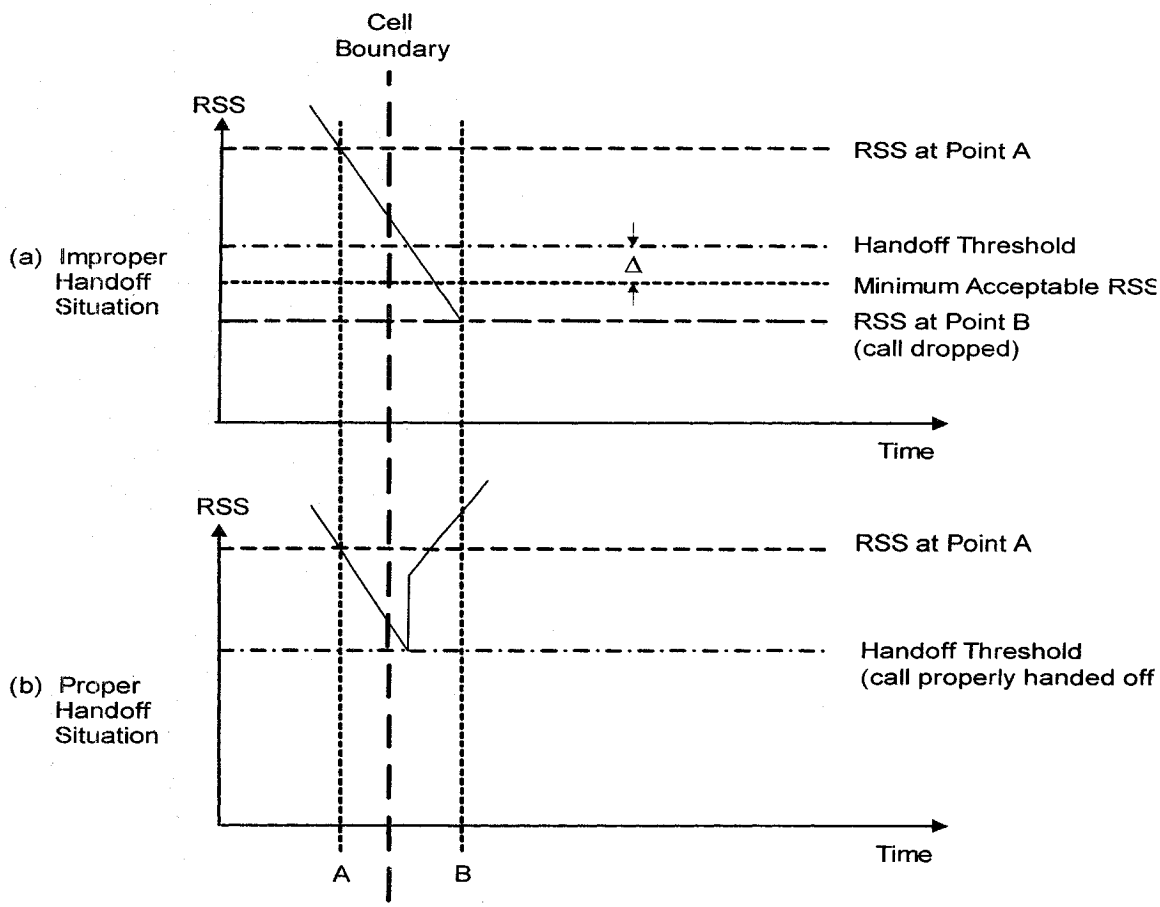

Figure 3.1: Handoff Scenario at Cell Boundary

Figure 3.1 (a) demonstrates the case where a handoff is not made and the signal drops below the minimum acceptable level to keep the channel active. This dropped call event can happen when there is an excessive delay by the core network element in assigning a handoff or when the threshold Δ is set too small for the handoff time in the system. Excessive delays may occur during high traffic conditions due to computational loading at the core network or due to the fact that no channels are available on any of the nearby base stations.

However, a major problem with this approach to handoff decision, based on signal strengths, is that the received signals of both base stations often fluctuate. When the mobile is between the base stations, the effect is to cause the mobile to wildly switch links with either of the base stations. The base stations bounce the link with the mobile back and forth. Hence the phenomenon is called *ping- ponging* [39] [45]. Besides the ping-pong effect, this approach allows too many handoffs [37]. Much of the time the previous link was well adequate and that handoffs occurred unnecessarily. A better method is to use the averaged signal levels relative to a threshold and hysteresis margin for handoff decision. The signal strength measures are signal levels averaged over a chosen amount of time. This averaging is necessary because of the Rayleigh fading nature of the environment in which the cellular network resides. Furthermore, the condition should be imposed that the target base station's signal level should be greater than that of the current base station.

3.1.1. Conventional Handoff Algorithm

The handoff algorithm employed in a cellular network has a significant impact on the overall network performance. It is to detect network situation and to decide whether it matches the handoff criterion. Under wireless circumstances, the RSS is a random process [39]. Some of the conventional decision algorithms [38] [39] [45] are as follows:

A. ***RSS***: The BS whose signal is being received with the largest strength is selected (Choose the new BS if $RSS_{new} > RSS_{old}$)

B. ***RSS plus Threshold***: A handoff is decided if the RSS of a new BS exceeds that of the current one, and the signal strength of the current BS is below a threshold T (Choose the new BS if $RSS_{new} > RSS_{old}$ and $RSS_{old} < T$)

C. ***RSS plus Hysteresis***: A handoff is decided if the RSS of a new BS is greater than that of the old BS by a Hysteresis margin H (Choose the new BS if $RSS_{new} > RSS_{old} + H$)

D. ***RSS, Hysteresis and Threshold***: A handoff is decided if the RSS of a new BS exceeds that of the current BS by a Hysteresis margin H, and the signal strength of the current BS is below a threshold T (Choose the new BS if $RSS_{new} > RSS_{old} + H$ and $RSS_{old} < T$)

E. ***Algorithm plus Dwell Timer***: Sometimes a dwell timer is used with the above algorithms. If the condition continues to be true until the timer expires, a handoff is performed [5].

3.2. Handoff in Heterogeneous Networks

It will be necessary for a MH to employ various points of attachments of a heterogeneous network to maintain connectivity to the network all the times. A new type of handoff, Vertical Handoff, has been defined [4] between base stations that are using different wireless network technologies. This is not the case of Horizontal Handoff, where both base stations use the same type of wireless network interface.

3.2.1. Different Mobility Architectures

We discuss five different mobility architectures [39] of heterogeneous network comprising WLAN and GPRS technologies as illustrated in Figure 3.2 for implementing handoff.

The ***first*** and ***second*** architectures involve connecting the WLAN and GPRS networks through GPRS entities such as the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). In these cases, the WLAN will appear to be a GPRS cell or Routing Area (RA), respectively. GPRS will be the *master network* and the WLAN will be the *slave network*. This means that the mobility will be handled by GPRS, considering the WLAN one of its cells or Routing Areas (RAs).

In the ***third*** scenario, Virtual Access Point (VAP) reverses the roles played by GPRS and WLAN in the first two architectures. Here, the WLAN is a *master network* and the GPRS is a *slave network*. The mobility is managed according to the IEEE 802.11 and Inter Access Point Protocol (IAPP) specifications by WLAN.

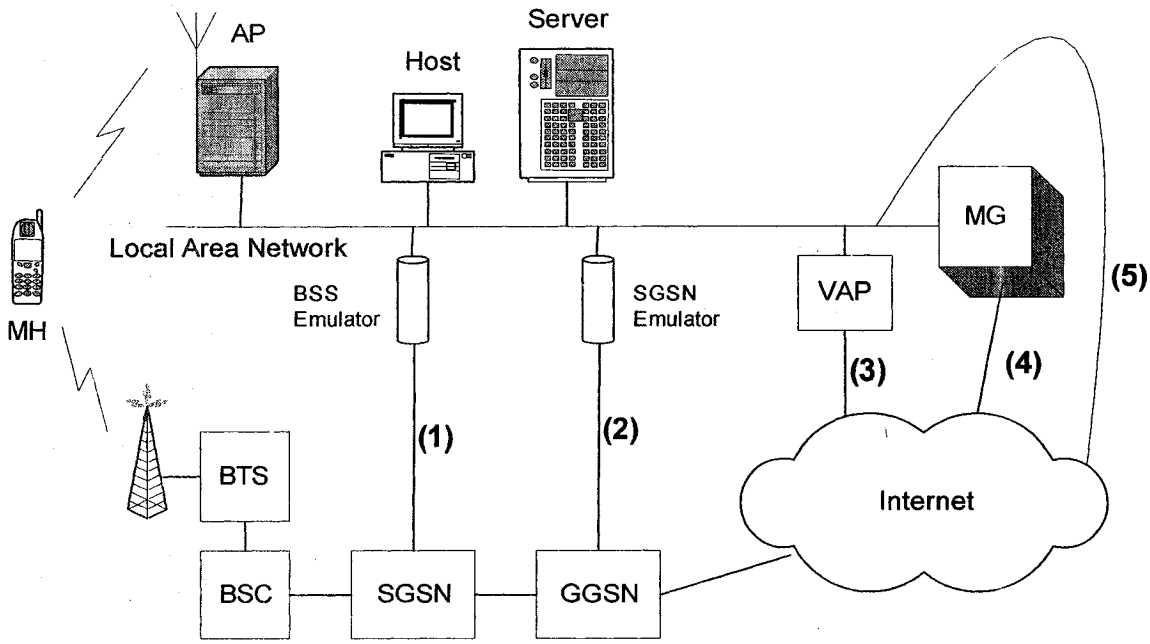

Figure 3.2: Different Mobility Architectures for Handoff between WLAN and GPRS

The ***fourth*** approach introduces a Mobility Gateway (MG) between the GPRS and WLAN network. The MG is a proxy implemented on either the GPRS or the WLAN sides, and will handle the mobility and routing issues.

The ***fifth*** architecture employs Mobile IP to handle the issue of mobility management. Here, GPRS and WLAN are peer networks.

3.2.2. Our Proposed Architecture

The first three architectures render one of the two networks as a slave network. In these cases, traffic of the slave network will occupy a considerable capacity of the master network, thus making it inefficient. The fourth architecture is not standardized, requiring proprietary protocols for inter-technology roaming [39]. Besides, the performance of a proxy is poor since significance latency is added to the client-server communication path.

The last architecture treats both the networks (i.e. GPRS and WLAN) as peer networks, and implements the function of Mobile IP respectively. Both the AP and BS is connected to the Internet using wireline or wireless connections. All the signals involved in the handoff process flow through the Internet. We propose the fifth architecture, the Mobile IP-based one, be used for managing the handoff between the two heterogeneous networks.

3.2.2.1. The Reason for Choosing Mobile IP-Based Architecture

We choose the Mobile IP-based handoff architecture as we consider it to be the appropriate tool to achieve the mobility between the two networks mainly for the following reasons:

1. The telecom industry has experienced an explosive growth of the Internet during the last half-decade or so. As the representatives of wireless wide area and local area data communication networks respectively, the GPRS and the WLAN will be widespread in near future. MH can easily access Internet through each of them.

2. An equally important factor is the growing use of the TCP/IP suite. TCP/IP is not just the suite for the Internet, but it has also recently become the most widely used data communications protocol to the desktop. Furthermore, looking towards the future, we can expect IP to be the pervasive communication protocol across a diverse set of devices, not just limited to computers. And Mobile IP is the logical evolution of TCP/IP.
3. Only the participating components (i.e. the MH and the endpoints of the tunnel) need to be Mobile IP aware. No other device in the network or any hosts with which the MH is communicating need to be changed or even aware of the movement of the MH.
4. It is secure because the set up of packet redirection is authenticated. The Registration Request, which is the integral part of Mobile IP-based handoff, uses unforgeable digital signatures for security purpose. Moreover, Mobile IPv6 will ensure even stronger authentication and encryption features.
5. Due to medium-independent design philosophy, Mobile IP does not require specialized hardware. Therefore, the implementation will be less costly than other alternative approaches
6. Expensive circuit-switched devices such as Mobile Switching Center (MSC), Home/Visitor Location Registers (HLR/VLR), Signal Transfer Point (STP) will replaced by off-the-shelf routers [8]. The technological trend to move towards all-IP-based interconnected network places Mobile IP as the number one choice for mobility management.

3.3. Issues Related to Handoff using Mobile IP

3.3.1. Handoff Latency

One of the important QoS parameters of our study in implementing the handoff using Mobile IP is latency. The Handoff Latency is the amount of time from when the MH is disconnected from the old AP/BS to the time when the mobile receives the first data packet from the new BS/AP [4]. The latency can be broken down into following components: (i) Discovery Time, (ii) Registration Time, and (iii) Packet Reception Time.

There is another factor in the overall handoff latency contributed to the time required for a MH to complete a handoff process at the Link Layer. This time for notification of Layer 2 to Layer 3 is very short. This value for a typical WLAN is usually in the range of several hundred microseconds up to several milliseconds [6].

3.3.1.1. Discovery Time

Discovering Time is the time during which the MH discovers that it has moved to a new Foreign Agent (FA) coverage area. It depends on the time required for several signal flow as below.

3.3.1.1.1. Beacon Period:

The AP/BSs send *beacon* periodically within their coverage areas. The MH receives these beacons from nearby AP/BSs and measures their strengths. Based on these measurements, the MH determines to which AP/BS it is likely to handoff in the future

[40]. Beacon Period is the time required by the MH to receive the next beacon signal from the AP/BS.

3.3.1.1.2. Agent Advertisement and Associated Messages:

Each AP/BS is associated with a FA, which is periodically transmitting *Agent Advertisement*. This period of time includes the time to receive a new advertisement after crossing a cell boundary. Internet Engineering Task Force (IETF) RFC 2002 has defined [25] the interval between two consecutive *Agent Advertising* messages to be 1/3 of the *Lifetime* [46], namely t. So, the Agent Advertisement Lifetime is 3t. IETF RFC 2002 also proposes t as 1 second.

The interval between two consecutive Agent Advertising messages as well as Lifetime is too long for a voice (real time) call to handoff. Therefore, the MH should actively send *Agent Advertisement Solicitation* message to accelerate the movement detection. If the *Agent Advertisement* message sent by the agents includes *Prefix Lengths Extension*, MH can decide whether it has changed subnet by comparing the subnet prefixes immediately after receiving one Agent Advertisement message. If so, it does handoff immediately, otherwise it does nothing.

There is a considerable amount of time for transmitting the Agent Advertisement and the associated messages. Following is the length of the messages to be used in our proposed handoff technique:

(a) Agent Advertisement Solicitation = 8 bytes

(b) Agent Advertisement (assuming 5 addresses) = 48 bytes

(c) Mobility Agent Advertisement Extension = 12 bytes

(d) Prefix-Lengths Extension = 4 bytes

For non real-time services, MH may wait for the periodic advertisement, instead of sending out *Agent Solicitation.*

3.3.1.1.3. Processing Time:

Because the length of Agent Advertisement message is very short, and there is only one hop between MH and the FA, the response will be quickly received. Generally, it takes agents about several milliseconds to process a request.

3.3.1.2. Registration Time

Registration Time is the time to complete a Mobile IP Registration process. In other words, it is the time to establish the new route to the MH. Followings are the main components of Registration Time:

Registration Request to FA → fixed value (very small, negligible)

Registration Ready from FA → fixed value (very small, negligible)

Registration Request to HA → random internet delay

Registration Ready from HA → random internet delay

Binding Update from AP/BS to FA → fixed value (very small, negligible)

Binding Update from FA to FA → random internet delay

Binding Acknowledgement from FA to FA → random internet delay

Binding Update from FA to CH → random internet delay

Binding Acknowledgement from CH to FA → random internet delay

Packet Forwarding from FA_{OLD} to FA_{NEW} → random internet delay + Packet Transmission Time (depends on packet size)

Processing Time → fixed value (very small, negligible)

The total time to complete the registration process will vary according to the scheme adopted for recovery of the lost packet during the handoff latency.

3.3.1.3. Packet Reception Time

Packet Reception Time is the time to receive the first forwarded IP packet after the new route is established. This is the time required to transmit the first packet from the new FA to the MH through the new AP/BS. Packet Reception Time is mainly dependent on the following two factors: (i) the transmission time of the data packets being forwarded through the air interface, (ii) the time required by the MH to accumulate the packet received from the AP/BS and process for output.

3.3.2. Lost Packet Recovery

In most of the cellular networks, handoff decision is made by the base station and there is a pre-setup on the neighbor channels. Thus the base station has a chance to duplicate the data on the neighbor base station and avoid data loss. However, in Mobile IP, the handoff

decision in Layer 3 is made by MH. And the handoff decision at Layer 2 is made independently from Layer 3 [6]. When the signal strength goes down to an unacceptable level, and the MH gets disconnected from the old AP/BS, the mobility agent (e.g. FA) may not be aware of this event immediately and thus packets addressed to the MH are dropped. However, these packets can be recovered by adopting various schemes.

In general, the lost packets recovery schemes can be classified as either *reactive* or *proactive*. The packets recovery method described in standard Mobile IP protocol is an example of reactive type. We define it as *Unicasting*. However, there is a proactive scheme, devised by some researchers and defined as *Multicasting*, which assists the MH in determining that a handoff is imminent, and establishes packet flow to the target FA prior to the handoff event [40] [41].

Both schemes have tradeoffs. We will employ both schemes in our study to take advantage of each type according to the service requirement.

3.3.2.1. Unicasting

In Unicasting scheme, the old FA buffers any data packets it is forwarding to an MH to reduce data loss during a handoff. When a handoff occurs, the MH includes a handoff request in its registration, and the new FA in turn requests that the old (previous) FA handovers the buffered packets. To reduce duplicates, the MH buffers the identification and source address fields in the IP headers of the packets it receives. It also includes the identification and source address fields in the buffer handover request so that the old FA

does not need to transmit those packets that have already been received by the MH. The message and data flow involving Unicasting scheme to recover the data packets are shown Figure 3.3.

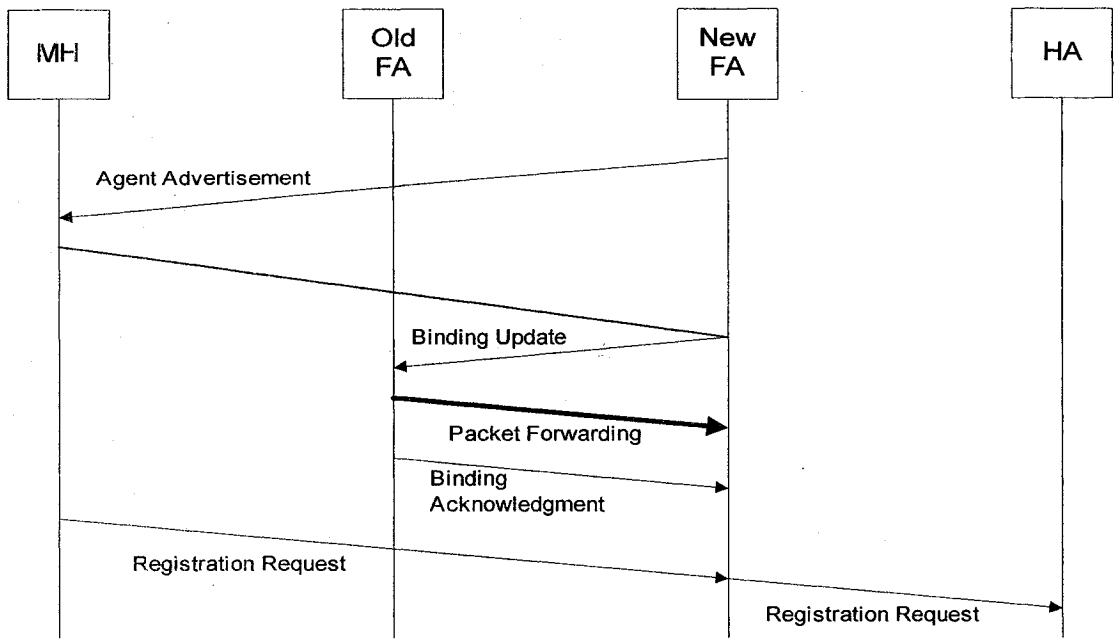

Figure 3.3: Unicasting Message and Data Flow

When the MH crosses a cell boundary, and it receives the first Agent Advertisement from the new FA, it sends a *Binding Update* message destined for the old FA via the new FA. The *Binding Update* message is routed by IP on a hop-by-hop basis to the old FA via the Internet. When this message is received at its destination (i.e. old FA), the packets are retransmitted along the new route just established to the new FA. Finally, the old FA sends an acknowledgement of the binding message destined for the new FA. This deletes the old entries along the old path.

3.3.2.2. Multicasting

The handoff latency can be reduced significantly if the network can set up data forwarding to the new FA while the MH is still communicating with the old FA. Today, this is possible in cellular networks, where the *soft handoff* procedure permits the MH to simultaneously receive signals from the old and new base stations. But this capability is neither available nor easily feasible in many current and emerging packet-based wireless networks. It is not desirable to impose such a capability as a requirement for Mobile IP, considering the complexity of predicting the new FA or the diversity of the wireless link technologies.

While it is difficult to know or predict the new FA exactly, it is not too difficult to find a reasonable set of candidate FAs (i.e. neighbors). One of those is likely to be the new FA after a handoff. If we allow proactive data forwarding (i.e. multicasting) to these prospective FAs just before, or at the very initiation of the handoff process, then we can achieve a lower handoff latency [41]. The scheme is built upon Mobile IP. The message and data flow involving Multicasting scheme to recover the data packets are shown in Figure 3.4.

As soon as the old FA gets a Handoff Notification message from a MH, it reads its Neighbor FA table, and sends Neighbor FAs the Forwarding Notification messages. Then the old FA starts forwarding data to the neighbor FAs. When the new FA receives data packets forwarded by the old FA the first time, the packets contain the home IP address

of the MH as the destination address, and the new FA needs to know the L2 address of the MH to forward the packets to the MH.

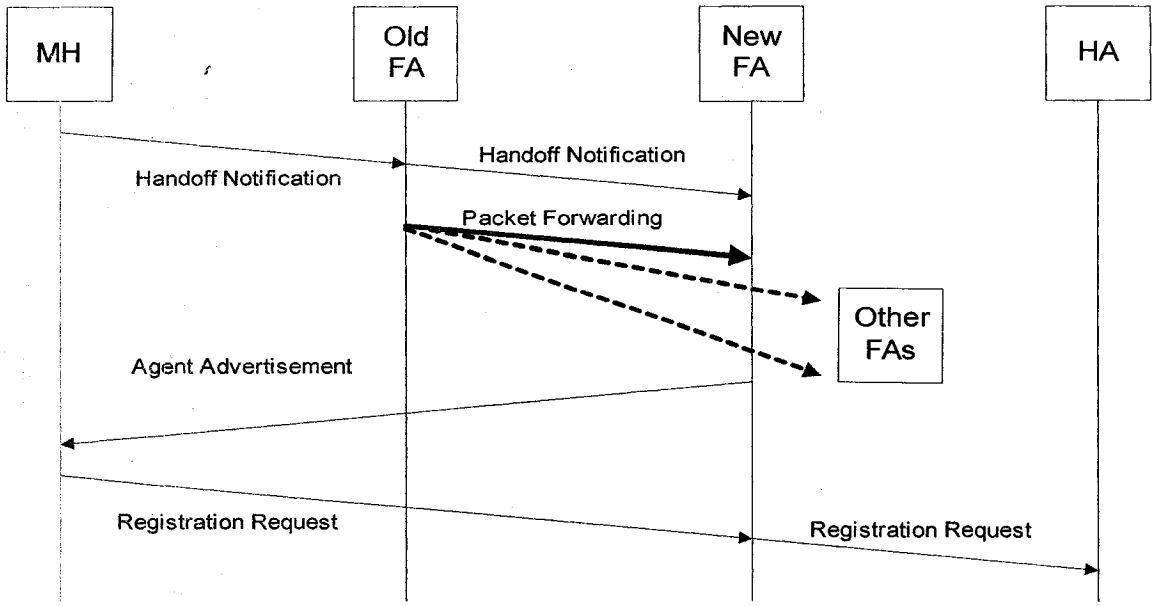

Figure 3.4: Multicasting Message and Data Flow

For IEEE 802.11 WLAN, the new FA can forward IP packets to the MH by relying on the MAC address of its interface. Since the MAC address is the same before and after the handoff, the old FA can provide the MAC addresses of the interface of the MH to the new FA, and the new FA can forward IP packets to the MH using provided MAC address. Then the new FA does not have to rely on the *Address Resolution Protocol (ARP)*. All the forwarded data packets to a neighbor FA are discarded if the FA is not the new FA. The key point is that the MH can receive data without finishing the Mobile IP registration process.

3.3.2.2.1. Neighbor FA Discovery Mechanism:

The discovery mechanism is executed in a distributed fashion; that is, there is no central controller or database. The size of the neighbor FA table is proportional to the number of adjacent or overlapping cells, which is practically bounded to be a moderate number. Each FA can learn about its neighbor FAs from the MHs since the MHs move around and come from the neighbor FAs.

The first step is that the MH includes the information about the old FA in the registration message sent to the new FA. Thus the new FA is informed of the old FA as one of its neighbor. Then it sends a Neighbor FA Notification message to the old FA. Each FA maintains a Neighbor FA table. The Neighbor FA information is of soft state; that is, each entry expires after the Neighbor Timeout Period has passed without any new notification for the Neighbor FA entry [41].

Chapter 4

4. Proposed Handoff Technique and Performance

In this chapter, we introduce a new handoff technique for handoff between WLAN and GPRS networks. The technique includes an algorithm to be implemented in the handoff decision making process as well as modification of the Mobile IP for handoff signaling management to achieve improved performance in latency, overhead, and throughput.

To improve the reliability of the channel quality measurements, we introduce two different signal strength measurements – Absolute Received Signal Strengths (ARSS) and Relative Received Signal Strengths (RRSS). Handoff is implemented if the ARSS or/and RRSS reaches predefined threshold value(s) based on the handoff direction (GPRS→WLAN, WLAN→WLAN or WLAN→GPRS).

Mobile IP-based handoff is associated with a latency which may not meet the QoS criteria for real time voice services. The data packets are lost during the latency period. Mobile IP describes a scheme, defined as Unicasting in our thesis, to recover the lost packets from the old FA to new FA. However, the process takes some time as the signal experiences a random delay when it travels through the Internet. This makes the latency even longer. In another data recovery scheme, namely Multicasting, data packets are sent to the neighboring FAs as soon as the Received Signal Strength (RSS) of the MH goes below a certain threshold level. The packets are stored in the buffer of the new FA, and in

the process, the latency can be reduced. We have simulated handoffs using the standard Mobile IP signals adopting both Unicasting and Multicasting schemes.

In this chapter, we also propose two modifications of the Mobile IP signal flows for further improvement of latency using Multicasting. Firstly, the new FA will not wait for the registration process to complete, and start sending packets to the MH after a fixed delay. Secondly, the packet size will be dynamically changing with the change of data rates. We have simulated handoffs using the standard Mobile IP and our proposed algorithm.

While Multicasting improves the latency, it also involves wastage of bandwidth, defined as Traffic Overhead, as data packets are being stored in neighboring FAs. We use Fluid Flow Model to measure the Traffic Overhead Ratio (TOHR). The problem of traffic overhead becomes very acute, and makes the proposition very inefficient, as the MH takes higher speed. To improve overall traffic overhead, we introduce a new approach in handoff decision making process. We define a new threshold type, namely *Distance Threshold*, which assumes two values – lower and higher - based on the values of ARSS and RRSS. However, the selection of the Distance Threshold value (lower or higher) will depend of the handoff direction as well as call type (voice or data). The MH will implement handoff as soon as the measured distance it travels in the handoff transition time goes below the appropriate Distance Threshold. In our thesis, we describe the selection process of the Distance Threshold.

4.1. Network Related Considerations of the New Handoff Technique

Our proposed handoff scheme has considered the following major network-related factors which make vertical handoff different from horizontal handoff:

1. Comparison of Received Signal Strength in heterogeneous networks is not as straightforward as in case of homogeneous network.
2. When two or more network interface technologies are integrated as peer networks to form a heterogeneous network, there is no core network elements (e.g. MSC, BSC) to coordinate the handoff. So make-before-break type of approach is not feasible.
3. Because of different Physical and Data Link Layers of heterogeneous networks (e.g. comprising WLAN and GPRS), the handoff is required to be managed at the Network Layer.
4. WLAN and GPRS cells constitute underlay and overlay networks respectively.

4.2. Key Characteristics of the Proposed Handoff Technique

1. Employs Mobile IP to handle the issue of handoff/mobility management. Here, GPRS and WLAN are peer networks.
2. Adopts two types of measurements of signal strength in the handoff decision making process for improved QoS.
    a. Absolute Received Signal Strength (ARSS) plus hysteresis
    b. Relative Received Signal Strength (ARSS) plus hysteresis

3. Employs Mobile-Controlled Handoff as the measurement information available to make a handoff decision is minimal.
4. The two subject networks support different data rates. WLAN can support a data rate of several Mbit/s, while GPRS can only supply tens to hundreds of kbit/s. So, a mobile host should utilize the high data rate of WLANs as much as possible.
5. Takes advantage of the reliability of the GPRS network, particularly for voice (real-time) communication.
6. Minimizes the handoff latency time, especially for voice (real time) communications.
7. Recovers the lost packets during the handoff latency period.
8. The mobile host requires dual mode (WLAN and GPRS) capability.

4.3. Proposed Handoff Interconnection Architecture

The proposed IP-based interconnection architecture using Mobile IP is shown in Figure 4.1. Followings are the network parameters and assumptions used in our handoff technique:

1. WLAN cells, which constitute the underlay network, are smaller in size. On the other hand, each GPRS cell, which constitutes the overlay network, covers area of several WLAN cells.
2. The Home Agent (HA), the Foreign Agents (FAs) and the Correspondent Host (CH) are interconnected through the Internet.
3. AP/BS and the corresponding FA may be collocated.

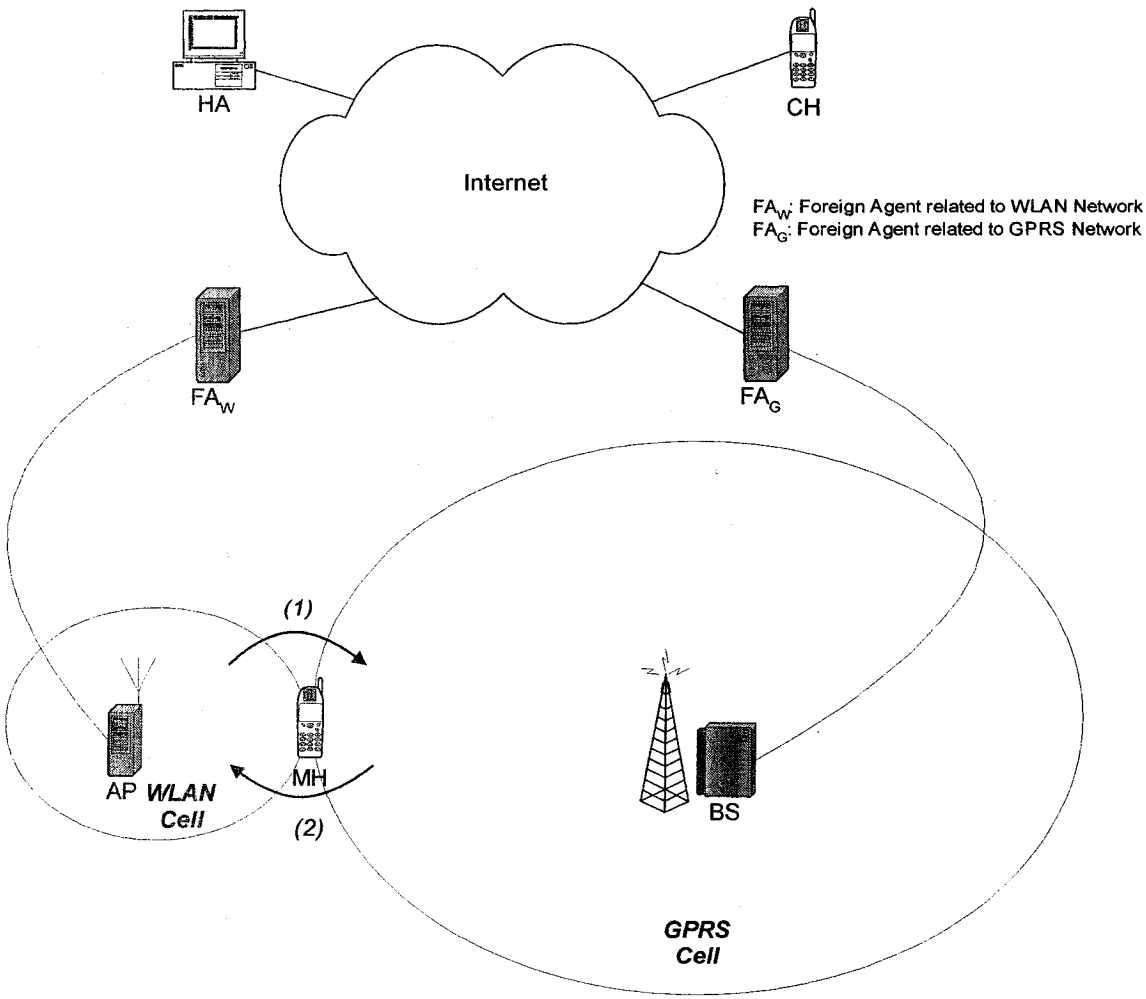

Figure 4.1: Proposed IP-based Handoff Architecture

4. FAs are connected to the Internet through wireless or wireless medium with very high bandwidth.
5. CH can be a fixed or mobile host.

4.4. Proposed Handoff Decision Algorithm

The schematic diagram depicting the general signal flow of the proposed algorithm is illustrated in Figure 4.2.

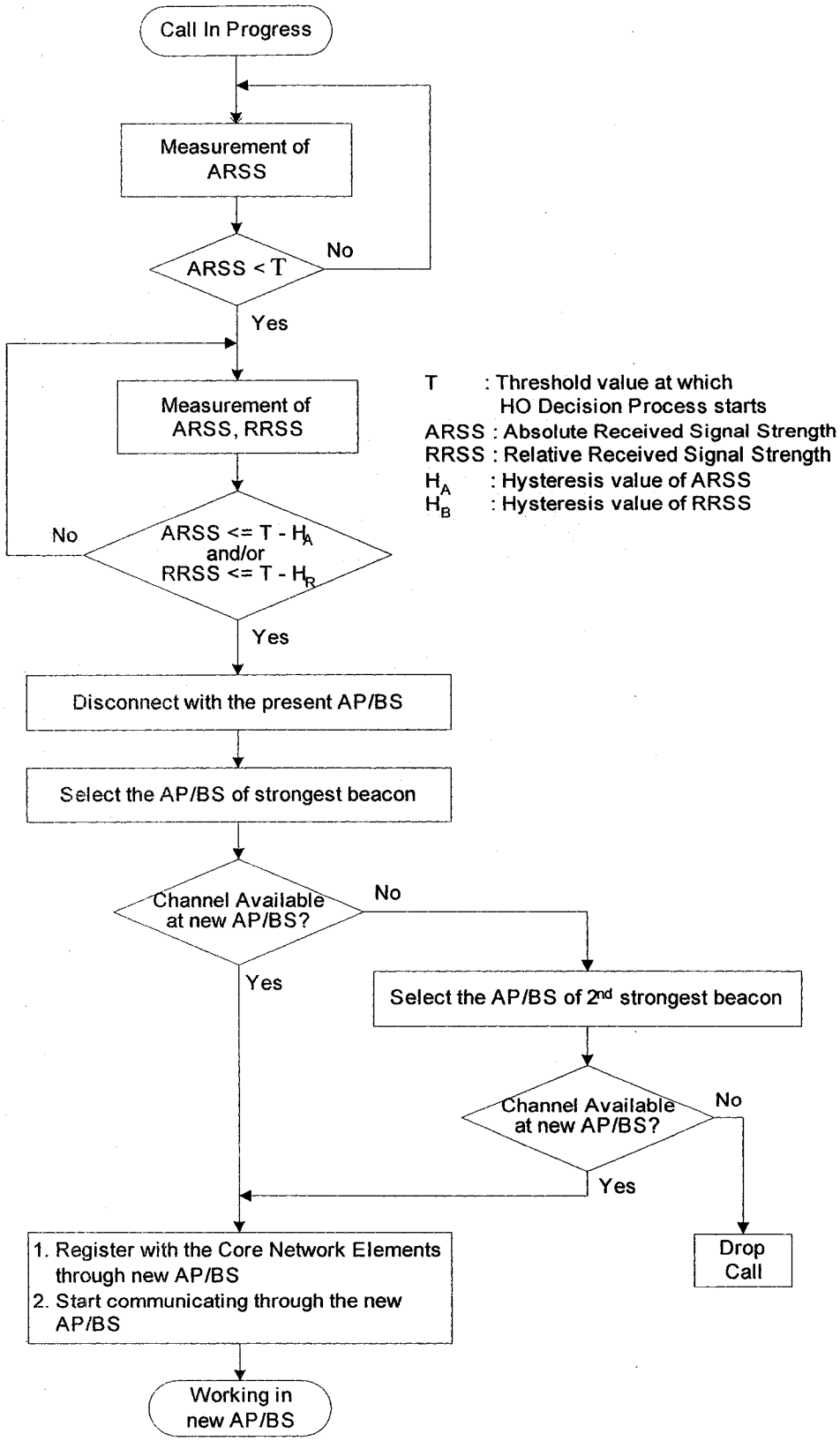

Figure 4.2: Proposed Handoff Algorithm

The MH periodically measures the Absolute Received Signal Strength (ARSS) of the AP/BS it is associated with, while the call is in progress. As the MH moves out of the coverage area of the current AP/BS, the ARSS decreases. The MH will take the samples of the ARSS from the AP/BS, and compare it with a predefined threshold T. When the value of ARSS falls below T, the handoff decision process starts.

The MH continually measures two signal values – (i) Absolute Received Signal Strength (ARSS) and (ii) Relative Received Signal Strength (RRSS) with the help of the beacon signals it periodically receives from its own as well as the neighboring AP/BSs. When ARSS and/or RRSS fall below predefined Hysteresis values H_A and/or H_B respectively, the MH is disconnected from the current AP/BS.

Now, the MH selects the AP/BS of the strongest beacon to be associated with. If the channel capacity is not available in the first selection to accommodate the handoff call, the MH looks for the AP/BS of the second strongest beacon. The call is dropped if the second AP/BS is also congested.

After selecting the new AP/BS, the MH gets registered with the core network elements through this new point of attachment. The MH starts communication through the new AP/BS after the registration is completed.

4.4.1. Signal Strength Measurement:

In our simulation, we are considering the distance from the AP/BS as the only factor for the Received Signal Strength (RSS). That means, the RSS increases as the MH gets closer to the AP/BS.

The free space signal power received by a receiver antenna, which is separated from a radiating transmitter antenna by a distance d, is given by,

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \qquad (4.1)$$

where, Pt = Transmitted Power

Gt = Transmitter Antenna Gain

Gr = Receiver Antenna Gain

λ = Wavelenth

L = System loss factor (not related to propagation)

Assuming all other variables except d for the MH to be same for both WLAN and GPRS environments, the power is inversely proportional to d^2. Using the above, we derive two different sets of signal strength measurements to be applied to our subsequent handoff algorithm.

4.4.1.1. Absolute Received Signal Strength (ARSS) Ratio

We consider that the actual handoff takes place at a point where,

P_l = MH Signal Power (Strength) at HO Execution Point due to departing AP/BS

d_l = Distance between the HO Execution Point and the departing AP/BS

P_2 = MH Signal Power (Strength) at the edge (boundary) of the departing AP/BS

d_2 = Radius of AP/BS

We define the ratio of the two signal power (strength) at the point, as illustrated in Figure 4.3, where the actual handoff takes place as follows:

$$\text{ARSS Ratio} = \frac{P_1}{P_2} = \frac{{d_2}^2}{{d_1}^2} \qquad (4.2)$$

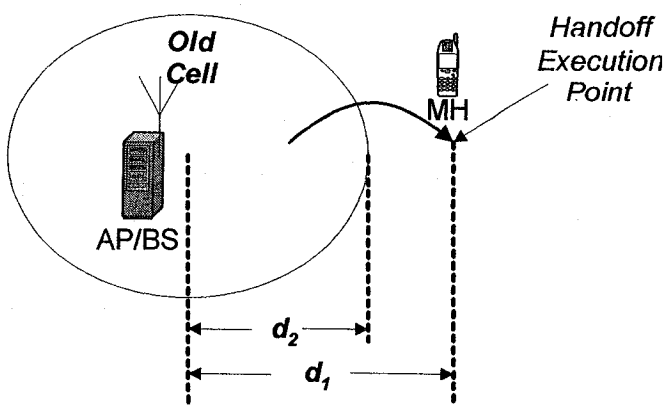

Figure 4.3: ARSS Measurement

4.4.1.2. Relative Received Signal Strength (RRSS) Ratio

We consider that the actual handoff takes place at a point where,

P_1 = MH Signal Power (Strength) at HO Execution Point due to old AP/BS

d_1 = Distance between the HO Execution Point and the old AP/BS

P_3 = MH Signal Power (Strength) at HO Execution Point due to new AP/BS

d_3 = Distance between the HO Execution Point and the new AP/BS

We define the ratio of the two signal power (strength) at the point, as illustrated in Figure 4.4, where the actual handoff takes place as follows:

$$\text{RRSS Ratio} = \frac{P_1}{P_3} = \frac{{d_3}^2}{{d_1}^2} \qquad (4.3)$$

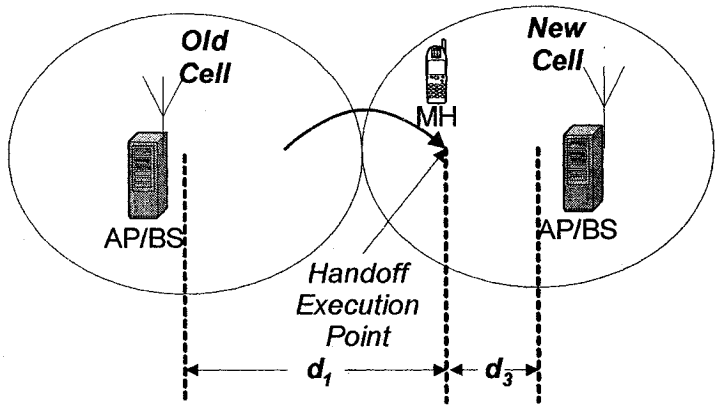

Figure 4.4: RRSS measurement

4.5. Mobile IP Signals in Proposed Handoff Implementation

Our proposed scheme implements Mobile IP to accomplish the handoff at the Network Layer. However, the Mobile IP handoff is executed only after a Link Layer handoff. A Link Layer handoff occurs when a MH changes from an AP/BS to another (they may belong to the same or different subnet). However, the Link Layer handoff time is very short, for instance, under the situation of wireless link, it is about 10 ms [46].

We have studied the two different schemes for recovery of the lost data packets during handoff latency period. Figure 4.5 and Figure 4.6 illustrate the overall signal flows in Mobile IP handoff while implementing Unicasting and Multicasting schemes respectively.

As illustrated in the figures, the difference in handoff latency between Unicasting and Multicasting scheme is because of the difference in Registration Time, while Discovery Time and Packet Reception Time are same.

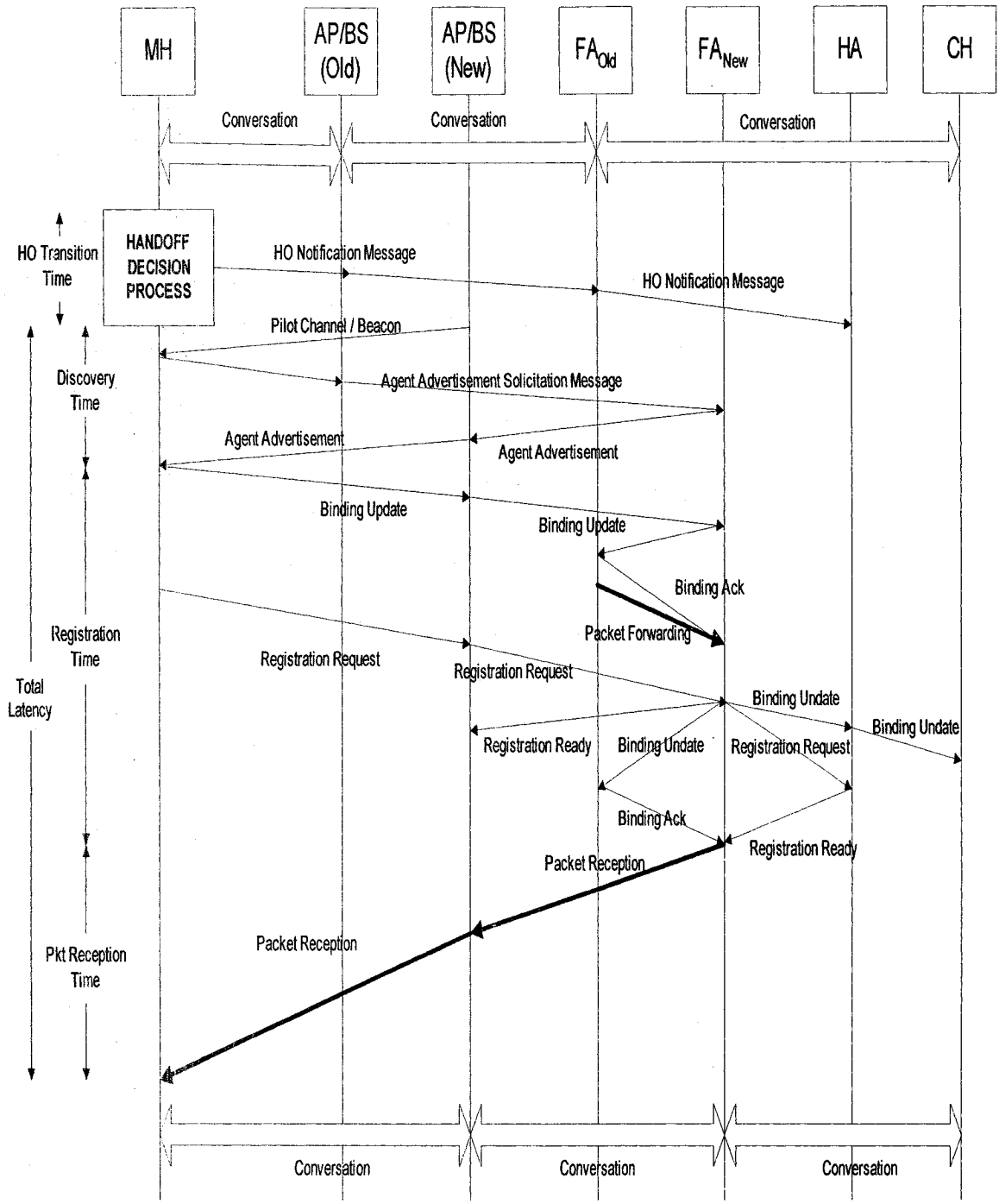

Figure 4.5: Signal Flow in Mobile IP-based Handoff - Unicasting

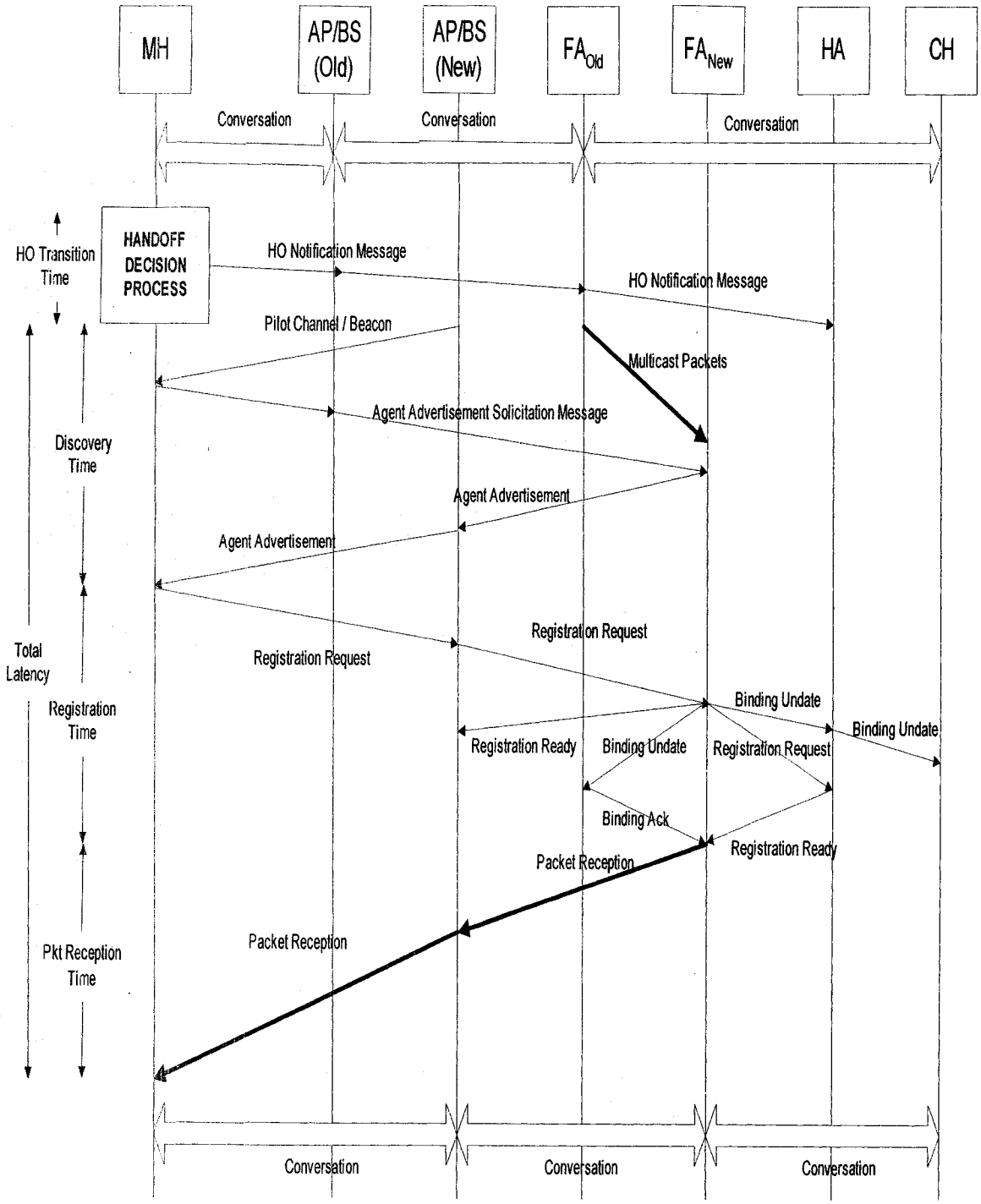

Figure 4.6: Signal Flow in Mobile IP-based Handoff - Multicasting

Just before the Layer 2 handoff (i.e. disconnection from the old AP/BS), the MH send a *Handoff Notification* message to the old FA and the HA. The MH waits for the beacon

from the nearby AP/BS of the strongest signal strength. After receiving the beacon, the MH sends *Agent Advertisement Solicitation* message to the new AP/BS through the new AP/BS, and receives the *Agent Advertisement*. In Multicasting scheme, the old FA starts forwarding the data packets to the new FA to store in the latter's buffer.

The next step is the registration which starts by sending *Registration Request* by the MH to the new FA through the new AP/BS. The new FA replies with *Registration Ready* message, forward the *Registration Request* to the HA, and updates the old FA and the CH by sending *Binding Update* messages. As it is evident from the figures, the Registration Time in Unicasting involves some additional flow of signals/data (e.g. *Binding Update*, *Binding Acknowledgement* etc.) between the two FAs. As such the total Handoff Latency for Unicasting would be higher than that of Multicasting.

After receiving the *Binding Acknowledgement* and *Registration Ready* messages from the old FA and the HA respectively, the new FA starts forwarding data packets to the MH through the new AP/BS.

We propose some improvements in the handoff scheme which is expected to reduce Registration Time and Packet Reception Time. Our proposed improvement scheme also focuses on minimizing the Overhead Ratio, which may become a QoS concern in different network situations in the Multicasting scheme.

4.5.1. Overhead During Handoff

Recovering the lost data packets is associated with traffic overhead as the traffic is forwarded to the neighboring FAs. This can be a significant for Multicasting scheme. This is an obvious cost of this scheme, while achieving low handoff latency as an advantage. On the other hand, Traffic Overhead is not significant in Unicasting applications.

To determine the overhead of traffic, we shall use a mobility model, which is suitable for our case. There are the two major kinds of mobility model which have been applied in previous location management studies. These are (1) Fluid-Flow Model, and (2) Random Walk Model [49].

Of these two models, Fluid-Flow Model is more suitable for users with high mobility and direction changes [52]. For pedestrian movements in which mobility is generally confined to a limited geographical area, such as residential and business buildings, Random-Walk Model is more appropriate [52]. In our thesis, we have considered the Fluid-Flow Model for analyzing the traffic overhead in the process of recovering the lost packets during handoff.

Under the Fluid-Flow Model, the direction of a MH's movement in a cell is uniformly distributed in every direction; i.e. in the range of $(0, 2\pi)$. According to the model, the rate of cell-boundary (or registration area) crossings; i.e. the handoff rate, can be shown [41] [50] [51] [52] as

$$H = (\rho . v . P) / \pi \quad (4.4)$$

where, H = Handoff Rate or Cell-Boundary Crossing Rate (1/sec)

ρ = Active Mobile Node Density ($1/m^2$); i.e. no. of user / m^2

v = Speed of the MH (m/sec)

P = Cell Perimeter (m)

Traffic Overhead Ratio, which is defined as the number of bits forwarded to the neighboring FAs divided by total number of bits sent by the previous FA to all the MHs in the cell, is

$$TOHR = D_{HO} / D_{TOTAL} \quad (4.5)$$

The amount of data forwarded from the previous FA to the neighboring FA(s) for a unit time period, $D_{HO} = m . T . R . H$ (4.6)

where, T = Period during which data is forwarded to FAs (sec)

m = Number of Neighboring FAs (where data is forwarded during time T)

R = Data Rate (bit/sec)

The total amount of data sent by the previous FA to all the MHs in the cell for a unit time period, $D_{TOTAL} = n . R$ (4.7)

where, n = Number of Active MHs in the cell

Furthermore, Active Mobile Node Density (ρ) can be defined as

$\rho = n / A$, where A = Cell Area (4.8)

Substituting the values of equation (4.4), (4.6), (4.7) and (4.8) in equation (4.5), we derive,

$$TOHR = (m.\ T.\ v.\ P) / (\pi.\ A) \quad \ldots\ldots (4.9)$$

The value of T in Multicasting is much higher in comparison with that in Unicasting. Besides, the value of m is several time higher in Multicasting than that (m=1) in Unicasting. Considering the above, it is evident from equation (4.10) that the Traffic Overhead will be much higher in Multicasting than in Unicasting. The overhead may become significant at higher speed of the MH. Our handoff scheme addresses this issue and proposes measures to reduce the overhead.

4.5.2. Throughput During Handoff

We define 'Throughput During Handoff' as the amount of data flow during the period of handoff decision making process. We measure the throughput in terms of Kilobit (Kb). During the Threshold Time, the MH communicates through the old AP/BS. Therefore, the throughput measured depends on the Data Rate and the Threshold Time of the MH while it is associated with the old AP/BS. Mathematically, we can define it as below:

$$\text{Throughput During Handoff} = R \times T \quad \ldots\ldots (4.10)$$

where, R = Data Rate

T = Threshold Time

4.6. Simulation and Performance Evaluation of the Proposed Handoff Technique

We have studied several Quality of Service (QoS) parameters of handoff between WLAN and GPRS networks using Mobile IP. We have proposed improvements of the handoff performance in the following areas:

- Handoff latency
- Overhead for lost packet recovery
- Throughput during handoff transition time

4.6.1. Simulation Model

Our model consists of an overlay network of GPRS with an under network of IEEE 802.11 WLAN. Our simulation environment consists of an underlay network of 3x3 WLAN cells, and an overlay network of 3x3 GPRS cells. However, the WLAN cells cover the area of the middle GPRS cell. The physical topology of the hybrid network for our simulation is shown in Figure 4.7. Followings are the main physical parameters as well as assumptions of our simulation model:

1. The radius of WLAN cells, r = 100m, 150m, 200m
2. Similarly, the radius of the GPRS cells, 3r = 300m, 450m, 600m
3. Number of AP per BS, n = 9
4. However, for simplicity, we have considered square-sized cell areas in our simulation.

5. The Access Points (AP) and the Base Stations (BS) are in the center of each cell. AP and BS are stationary. We represent the location and coverage area of the AP and BS by Cartesian coordinate system.

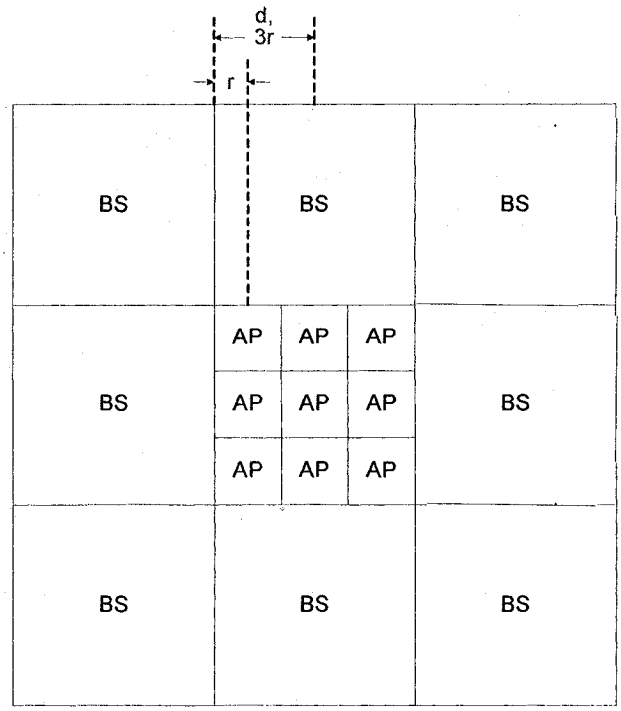

Figure 4.7: Physical Topology of the Simulation Model

4.6.1.1. Network Parameters and Assumptions

Initial Location of MH:

We randomize the initial locations of the users. Each user will start from a randomly selected location to be defined by its Cartesian coordinates $[X_i, Y_i]$.

Movement of MH:

Each user will start from its initial location to a randomized direction. Let the distance of user i be d_i and its angle (with the X-axis) be θ_i.

Here, θ_i is a random variable (varying between $0^0 \rightarrow 360^0$).

The distance along the X-axis is = $X_i \cos\theta_i$

The distance along the Y-axis is = $Y_i \sin\theta_i$

Speed of MH:

Each user will have a random speed varying between 1 m/s and 30 m/s; i.e. between 3.6 km/hour and 108 km/hour.

Change of Speed of MH:

The user changes its speed every 4 seconds.

***Check User Location*:**

The user location is checked in every 10 ms (i.e. 0.01 s). To update its location,

$X_i = X_i + V_i \times 0.01 \times \cos\theta_i$

$Y_i = Y_i + V_i \times 0.01 \times \sin\theta_i$

Call Generation:

When a user is created in our simulation grid, we consider a successful call has been generated. We treat it as an active user.

Internet Delay:

Internet Delay is a random variable in the range 0 ~ 95 ms.

***Data Rate*:**

Each user, at the time of generating a call, is associated with a data rate. This rate is randomly selected. However, we have assigned different ranges of data rates for GPRS and WLAN cells as follows:

Within GPRS Cell: Data Rate will be any of the following values, which are chosen randomly irrespective of voice or data calls → 8, 16, 24, 32, 40, 48, 56, 64 Kbps

Within WLAN Cell: Data Rate depends on voice or data calls. For voice calls, the rate is chosen randomly from any of the following values → 8, 16, 24, 32, 40, 48, 56, 64 Kbps. However, for data calls, the rate can take any random value in the range of 8~1000 Kbps.

Change of Data Rate:

From GPRS to WLAN Handoff: Data Rate is assumed to remain same.

From WLAN to GPRS Handoff: Data Rate is lowered, if required, to adapt with the GPRS capacity.

Beacon Period:

Beacon Period is random variable in the range 0 ~ 100 ms.

4.6.1.2. Simulation Program

We have developed our simulation program in C++ language. There are four basic modules of the program to simulate our network model. The modules are described briefly as follows:

(1) User:

- It's a class template
- It takes all local/global variable associated with mobile host; e.g. speed, data rates, probability of voice/data calls, lost packet recovery scheme etc.
- Once created, an object of user will be active in our main module
- The object will be deleted when the program terminates

(2) Cell:

- It's a class template

- It is used to form both WLAN and GPRS grids, within which a mobile host will be active
- Once created, an object of cell will be active in our main module
- The whole grid will be deleted when the program terminates

(3) Handoff:

- It's a class template
- It takes all local/global variable associated with handoff; e.g. beacon period, internet delay, threshold time etc.
- Once created, an object of handoff will be active in our main module
- The object will be deleted when the program terminates

(4) Main:

- It's the main module, which encapsulates all the other classes.
- This module is responsible for initiating, running to generate the results, and terminating the program (as shown in Figure 4.8)

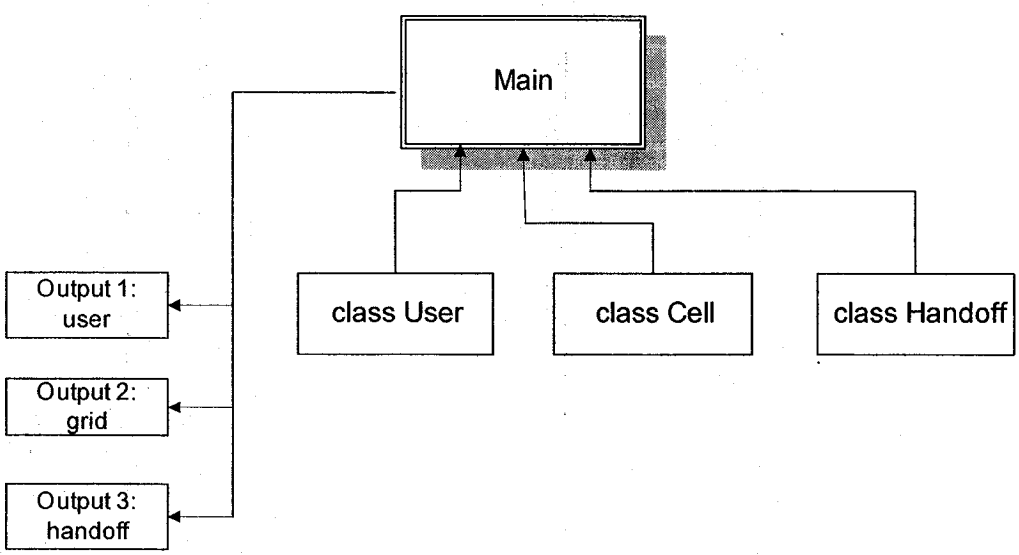

Figure 4.8: Module Representation of the Simulation Program

4.6.2. Handoff Latency Performance

4.6.2.1. Latency with Unicasting and Multicasting Using Standard Mobile IP

We have simulated the handoff in our model network to observe the distribution of handoff latency period for both Unicasting and Multicasting schemes (forwarding packets from the old FA to the new FA for lost recovery of the lost packets during handoff transition) with the following network parameters:

Beacon Period: Random value (0 ~ 100 ms)

Internet Delay: Random value (0 ~ 95 ms)

Data Rate: Random value (8 ~ 64 Kbps, at a multiple of 8)

Packet Size: Fixed (2 Kb)

Number of Handoff Studied: 224 (for each case)

Confidence Interval: For 95% confidence level, the number of handoffs we have considered, would give an error margin of 3.8% and 5.1% for Unicasting and Multicasting respectively

4.6.2.1.1. Unicasting:

We have obtained the following results with the distribution of handoff latency. Table 4.1 shows the latency ranges while Figure 4.9 shows the distribution.

Delay Range (ms)	*Percentage of Handoffs*
0 ~ 150	0%
151 ~ 400	73%
Above 400	27%

Table 4.1: Latency Ranges using Standard Mobile IP – Unicasting

Average Latency = 363 ms

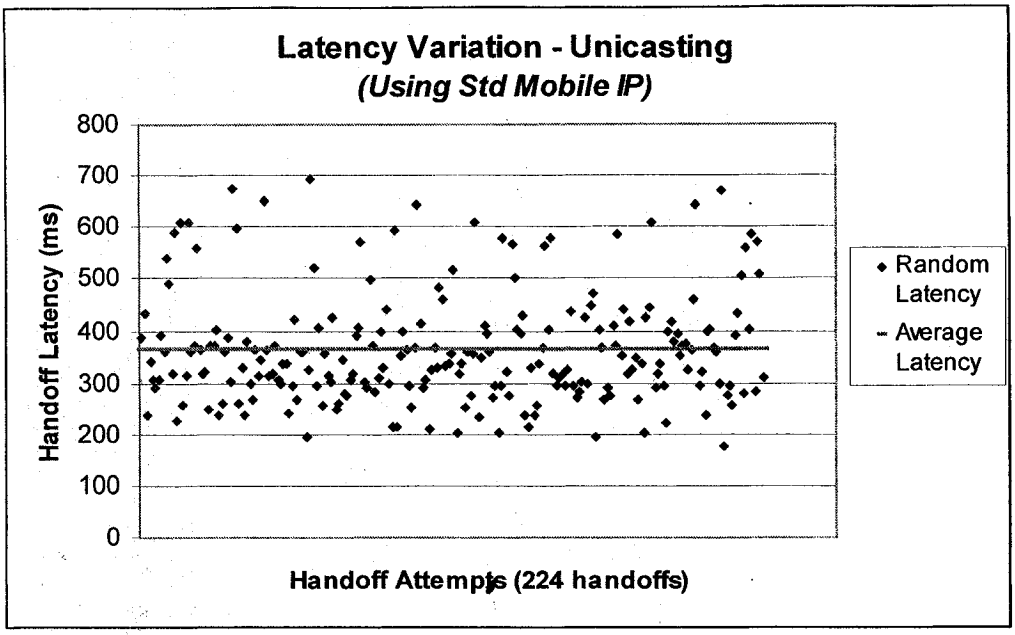

Figure 4.9: Latency Distribution Adopting Unicasting Scheme

4.6.2.1.2. Multicasting:

We have obtained the following results for the distribution of handoff latency. Table 4.2 shows the latency ranges while Figure 4.10 shows the distribution.

Delay Range (ms)	*Percentage of Handoffs*
0 ~ 150	7%
151 ~ 400	80%
Above 400	13%

Table 4.2: Latency Ranges using Standard Mobile IP – Multicasting

Average Latency = 264.33 ms

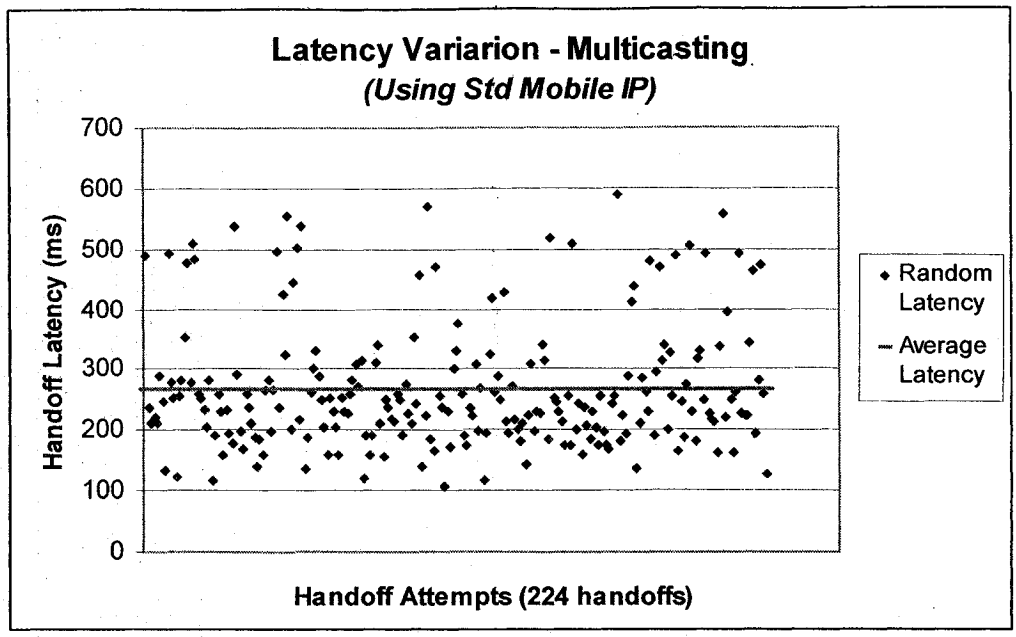

Figure 4.10: Latency Distribution Adopting Multicasting Scheme

4.6.2.1.3. Comparison of Latency for Unicasting and Multicasting:

The breakdown of the range of handoff latency as obtained from our simulation for the two schemes is illustrated in Figure 4.11.

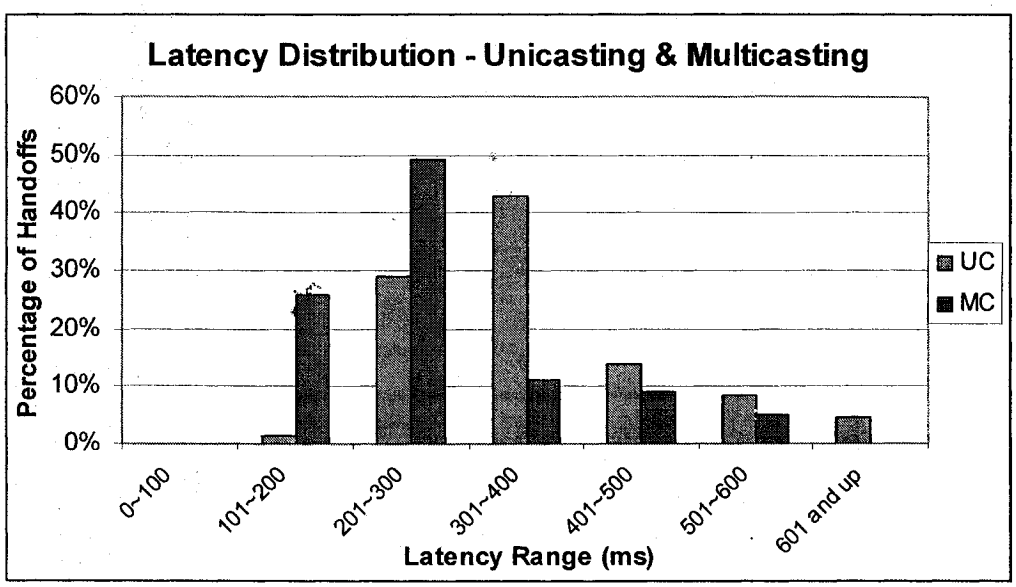

Figure 4.11: Comparison of Latency for Unicasting and Multicasting

From the distribution and average values obtained from the simulation we observe that the handoff latency for Unicasting is significantly higher than that for Multicasting. Moreover, a considerable percentage of handoffs using Unicasting require more than 400 ms to complete. For voice (real-time) communication, this is a significant amount of delay to maintain even a reasonable QoS. However, for data communication, the delay is not a large factor. Rather, loss of packets during handoff plays a key role in maintaining the QoS of data calls.

In view of our finding above, we propose Multicasting to be used for voice calls while Unicasting to be used for data calls as depicted in Figure 4.12.

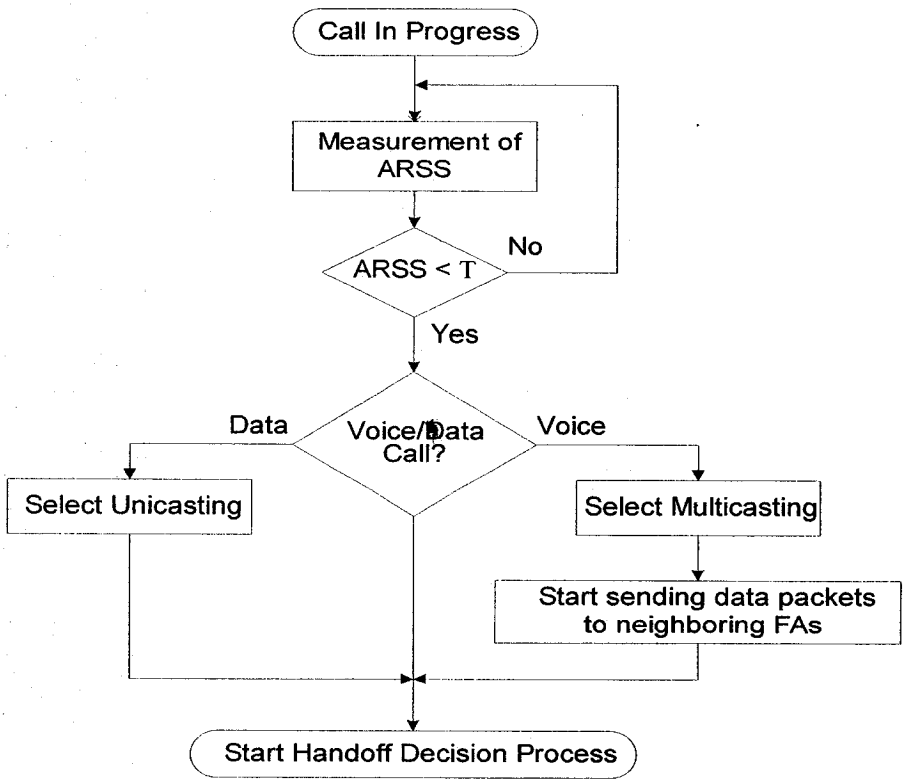

Figure 4.12: Selection of Lost Packet Recovery Scheme

4.6.2.2. Proposed Improvement in Latency

4.6.2.2.1. Improvement in Registration Time:

In this section, we propose an improvement in Registration Time by starting to forward the data packets after a small fixed delay (we term it as 'Fixed Registration Delay') following the Registration Request from the MH to the new FA through the new AP/BS. That means the data packet will not wait for the Registration process to complete. As the new FA has multicast data packets stored in its buffer, it can start sending those to the MH immediately after receiving the Registration Request from the MH. This will reduce the total handoff latency significantly. The process also reduces the requirement of high buffer capacity in the FA. The improvement mechanism is schematically illustrated in Figure 4.13. The solid line and the dotted line show our proposed signal flow and the original (Std Mobile IP) signal flow respectively.

We also propose that the new FA directly sends the Binding Update to the Correspondent Host (CH) instead of sending it to the HA to forward to the CH. Thus the CH will be notified about the new point of attachment of the MH earlier. This would help reduce the number of data packets forwarded to the previous FA even after the MH has registered with a new FA, and thus reduce the probability of packet loss during handoff.

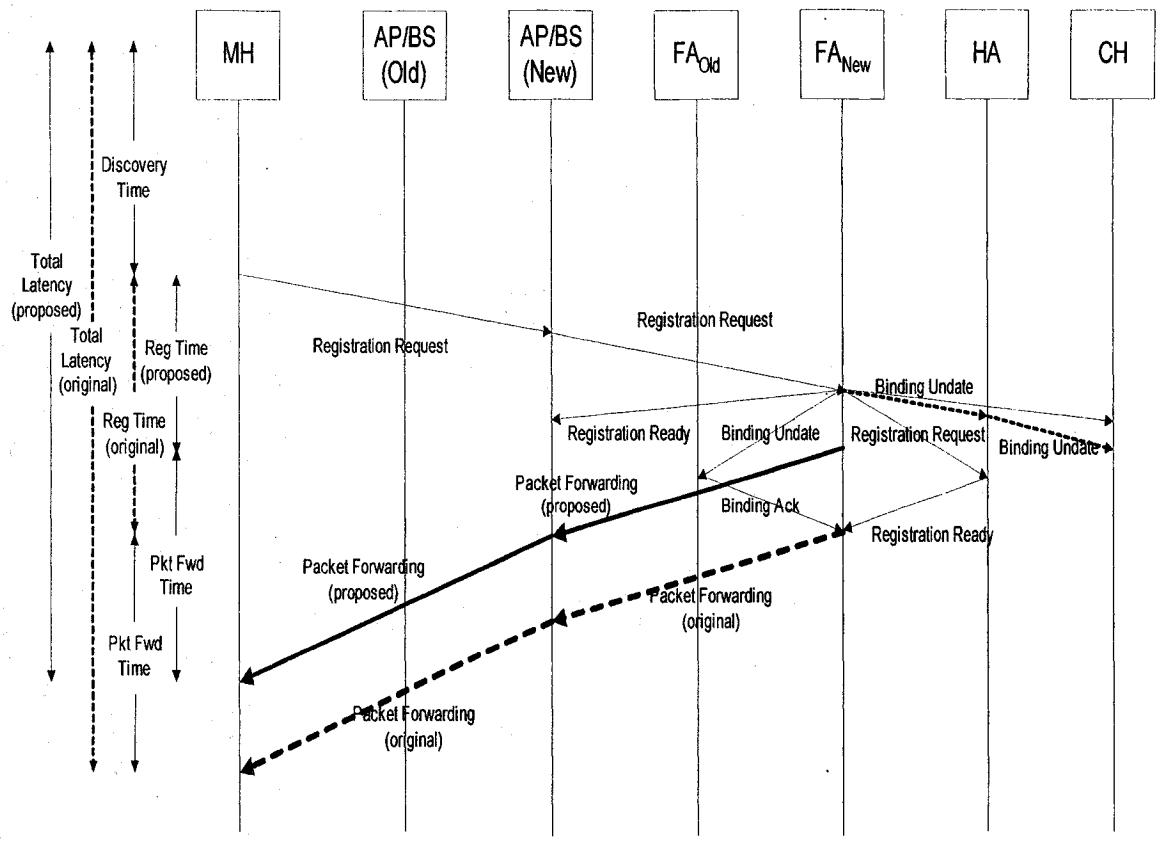

Figure 4.13: Improvement in Registration Time

While the handoff latency is reduced with the new technique as in Figure 4.13, there are still consequences in implementing this process. Firstly, there is a possibility that registration process with the HA becomes unsuccessful for some reasons (e.g. authentication). In this case, the call will be dropped after a fixed time period (we term it as 'Registration Lifetime', as depicted in Figure 4.14) if the new FA does not receive the Registration Ready message from the HA within that period. Secondly, there may be delay in Binding Update (BU) and Binding Acknowledgement (BA) messages to and from the previous FA. In that case, the call will be dropped after a fixed time period (we term it as 'Binding Lifetime') if the new FA does not receive the BA from the previous FA within that period. The implementation of the improvement algorithm is illustrated in Figure 4.14.

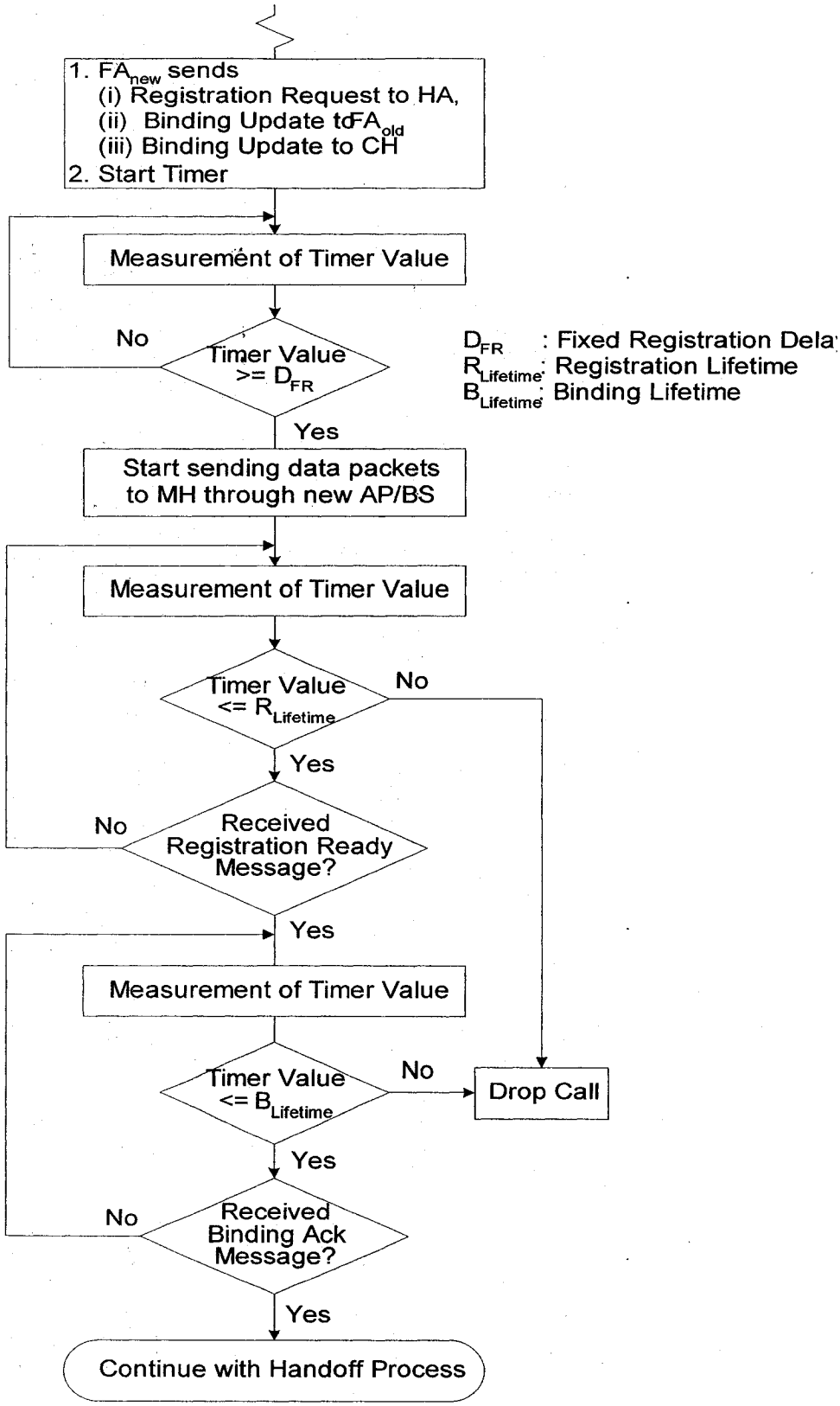

Figure 4.14: Implementation of Improvement in Registration Time

4.6.2.2.2. Improvement in Packet Reception Time:

The main contributors to the Packet Reception Time is the time required for transmitting the data packets to the MH and the time required by the MH to process to finally available to the user. It is dependent on the Packet Size and the Data Rate of the transmission channel. For a low Data Rate application, the transmission takes a significant amount of time, whereas it is not a major problem for higher rate applications. As most of the voice calls require low data rates, the Packet Reception Time is a substantial factor in the overall handoff latency.

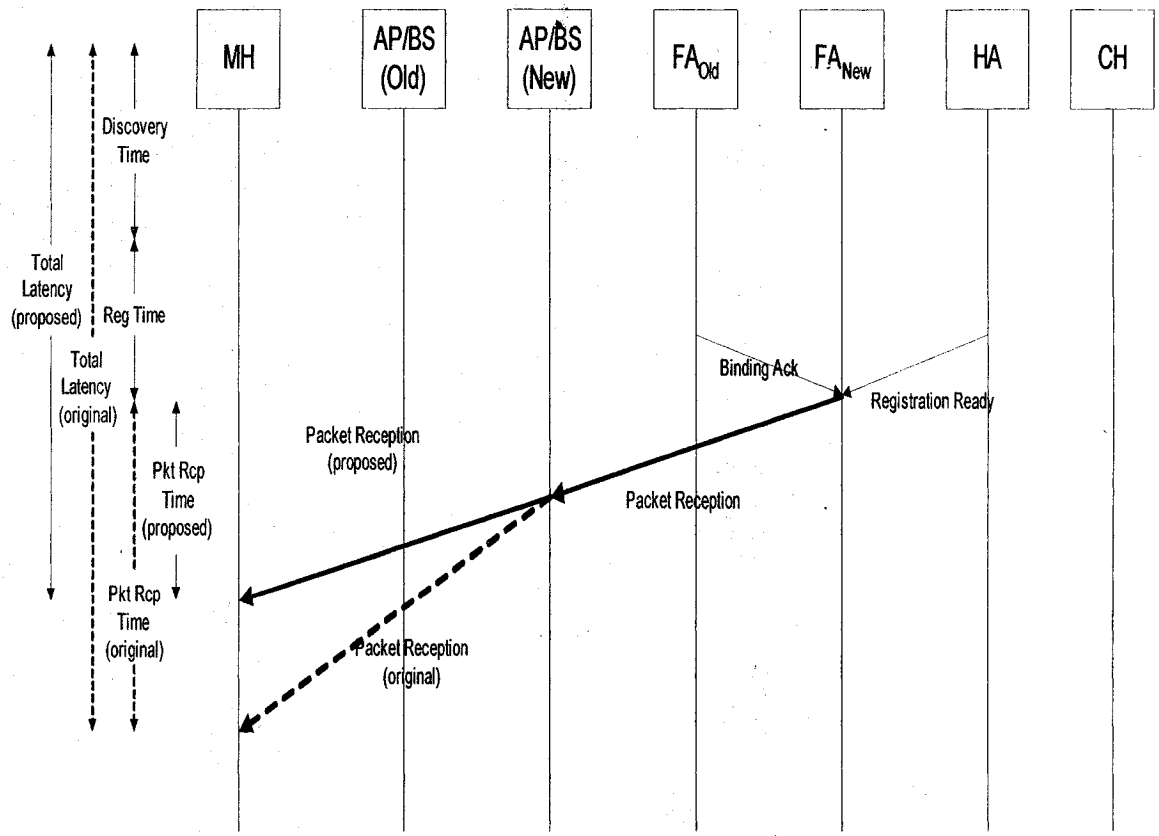

Figure 4.15: Improvement in Packet Reception Time

In this section, we propose an improvement in Packet Reception Time by implementing adaptive packet size. In our scheme, the network will adjust the packet size according to the data rate of the call. The packet size will be small (or large) in proportion to the data rate applications. The improvement mechanism is schematically shown in Figure 4.15.

There is a positive effect for a smaller packet size on the amount of packet losses. A smaller packet size makes the period of packet transmission smaller. In consequence, the duration of packet loss also gets smaller [6]. However, we have not studied the effect of packet loss in this thesis.

There is another implication on lowering the packet size. While we can improve the handoff latency as well as packet delay, it is associated with decreasing the transmission efficiency. As IP data packets are accompanied with a considerable header size, it can be very significant for a low packet size. However, with the implementing of header compression techniques, the problem can be greatly eliminated.

4.6.2.3. Simulation Results

We have simulated the handoff in our model network to observe the distribution of handoff latency period for both Unicasting and Multicasting schemes after implementing our proposed algorithm. We have kept the Beacon Period, Internet Delay, and Data Rate same as those used in our simulation of the Standard Mobile IP algorithm. However, we have used the Packet Size in the range of 1~2 Kb, which adapts proportionately to the data rate in the range of 8~64 Kbps. We have considered the packet size to change

proportional to the square of the data rate. We have used nominal value of 20 ms for Registration Lifetime. Figure 4.16 shows the distribution of handoff latency for 224 handoff cases being studied using the proposed algorithm. For 95% confidence interval, the 224 handoff samples provides a maximum error of 7.33 ms, which is 4.68%.

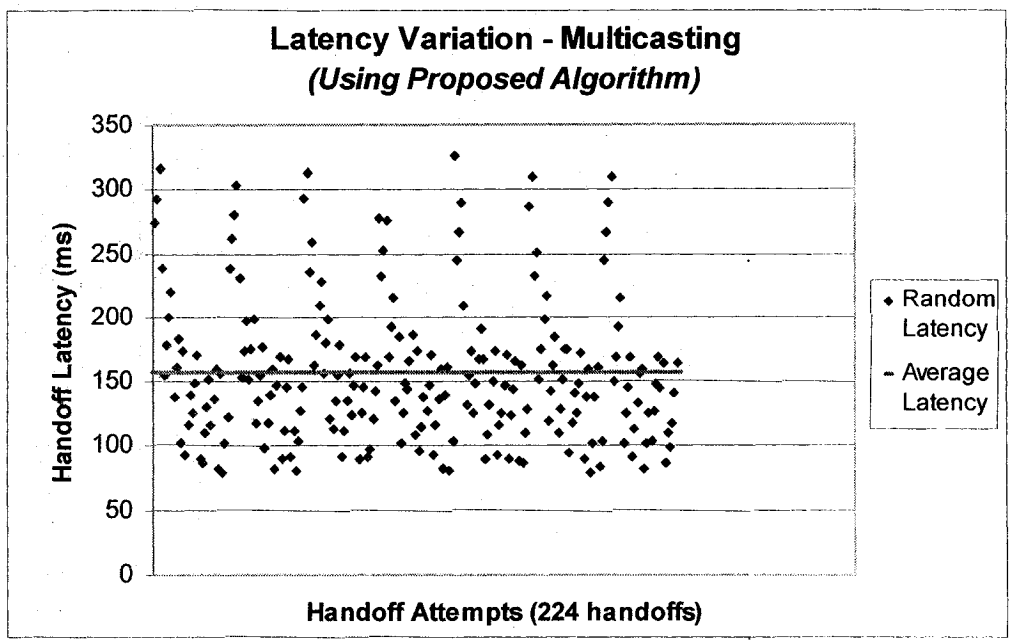

Figure 4.16: Latency Distribution Using the Proposed Algorithm

We have obtained the following results for the distribution of handoff latency as shown in Table 4.3:

Delay Range (ms)	Percentage of Handoffs
0 ~ 150	52.68%
151 ~ 400	47.32%
Above 400	0%

Table 4.3: Latency Ranges using proposed algorithm - Multicasting

Average Latency = 156.57 ms

4.6.2.3.1. Comparison of Latency for Standard Mobile IP and the Proposed Algorithm

From the distribution and average values obtained from the simulation we observe that the handoff latency using our proposed algorithm is significantly lower than that using the standard Mobile IP protocol. The comparison is illustrated in Figure 4.17.

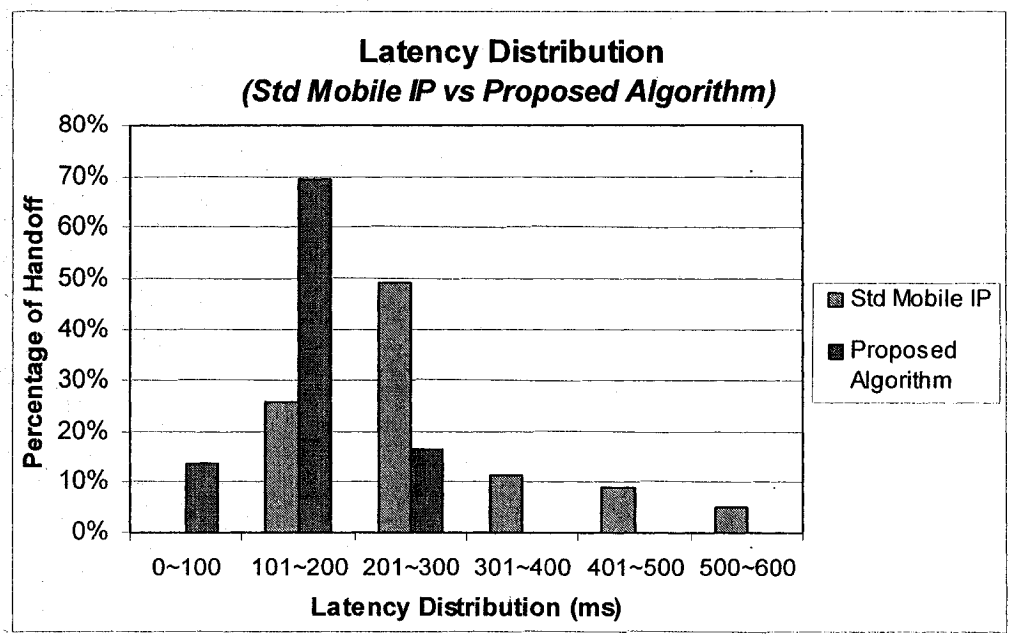

Figure 4.17: Comparison of Latency for Standard Mobile IP and the Proposed Algorithm

We have also simulated the handoff with varying the data rate while considering some fixed values of Beacon Period (50 ms) and Internet Delay (50 ms). Our results are shown in Figure 4.18, 4.19 and 4.20.

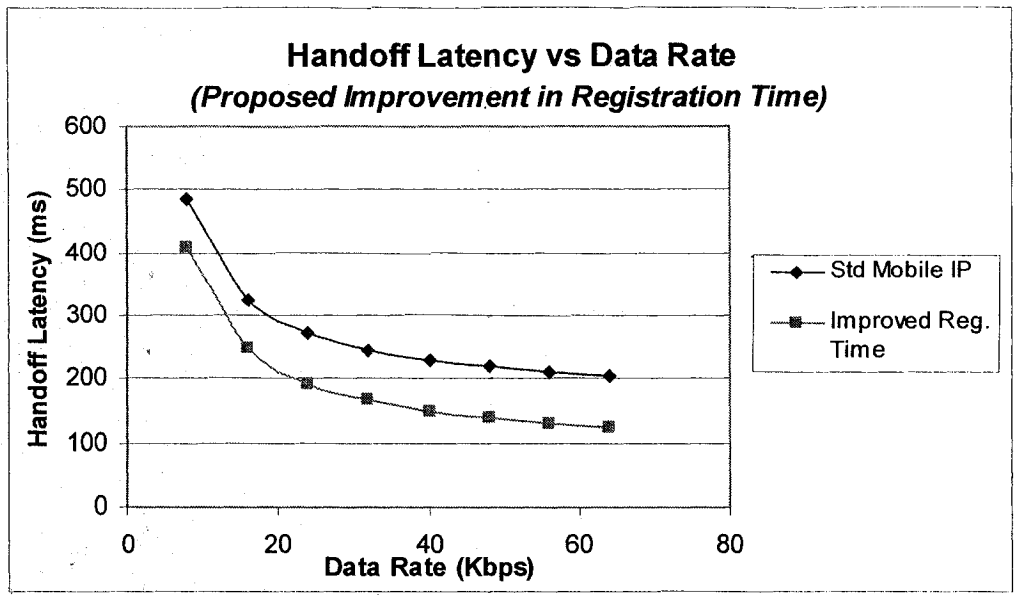

Figure 4.18: Effect of the proposed improvement in Registration Time on overall Handoff Latency with varying Data Rates

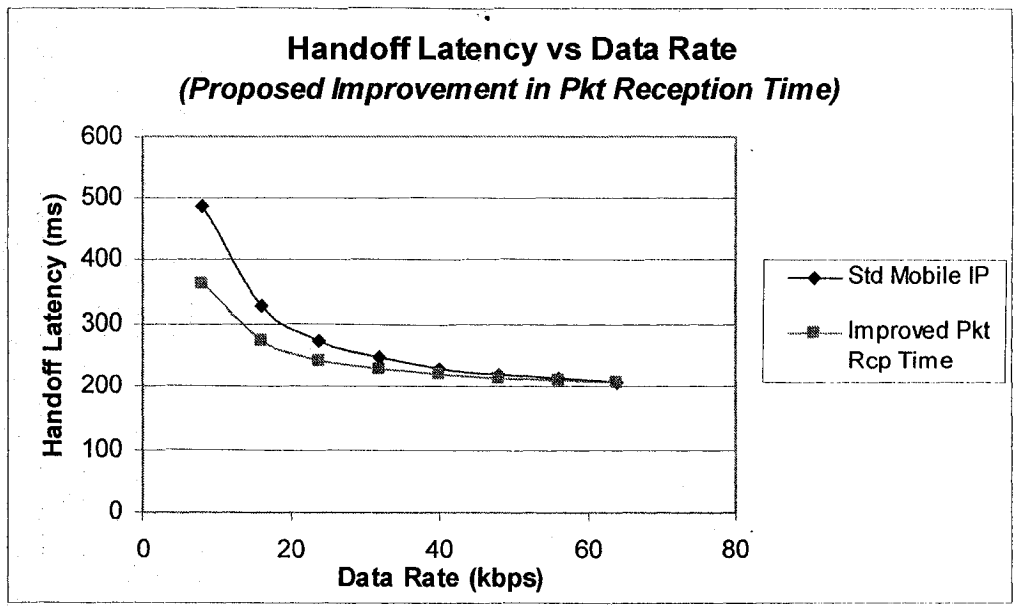

Figure 4.19: Effect of the proposed improvement in Packet Reception Time on overall Handoff Latency with varying Data Rates

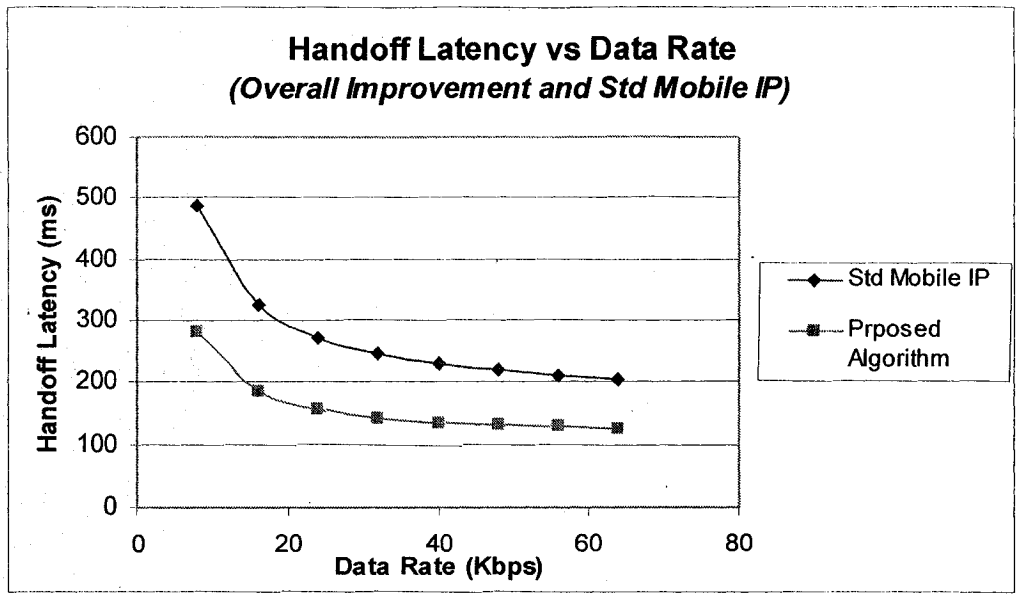

Figure 4.20: Comparison of Handoff Latency with varying Data Rates: Standard Mobile IP vs. Proposed Algorithm

The above results clearly show a significant improvement in the handoff latency. However, it can be achieved at the expense of some system complexity.

4.6.3. Traffic Overhead Performance

We have simulated handoff using Multicasting scheme for GPRS→WLAN, WLAN→WLAN and WLAN→GPRS, and studied the Traffic Overhead Ratio (TOHR) at different MH speed. Figure 4.21 shows the change of average TOHR for handoffs simulated with the following parameters:

Speed: 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 30 m/s

Threshold Time: 10 s

Number of handoffs simulated: 10 (in each speed) for GPRS→WLAN, 3 (in each speed) for WLAN→WLAN and 3 (in each speed) for WLAN→GPRS

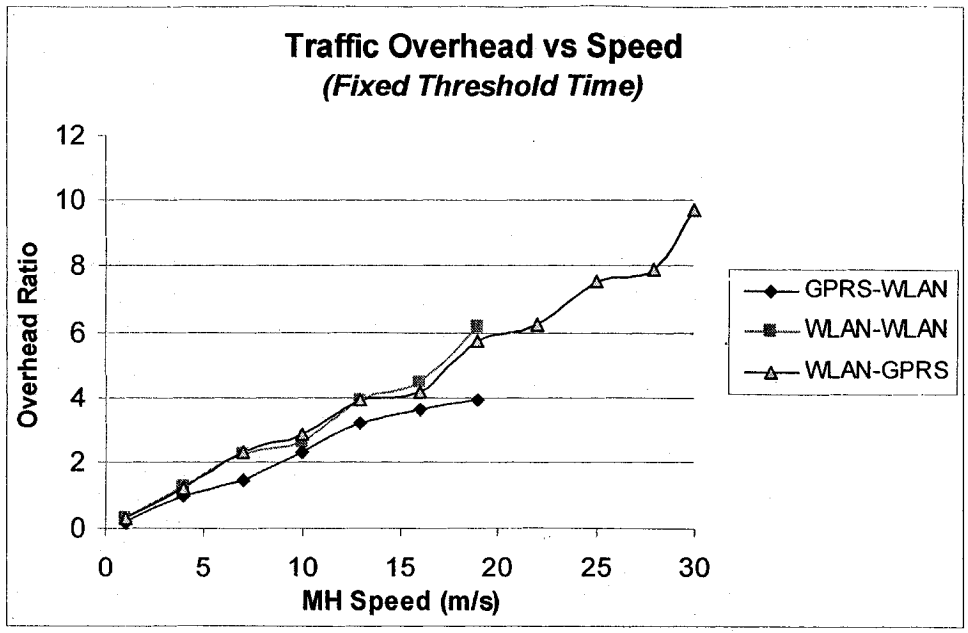

Figure 4.21: Traffic Overhead for various handoff scenario at varying speed with fixed Threshold Time

As we observe TOHR increases with the increase of MH speed. At lower speeds (around 1 m/s or 3.6 km/hour) the overhead may be acceptable. However, at higher speeds, the overhead increases to a high value that the Multicasting scheme will prove to be highly inefficient scheme for lost packet recovery.

The difference in values of TOHR for the various handoff scenarios is due to the fact that the number of FA (i.e. in this case the AP/BS) varies. TOHR takes more or less the same pattern for WLAN→WLAN and WLAN→GPRS as the number of neighboring FAs varies only a little in our simulation model. On the other hand, TOHR is less with GPRS→WLAN handoff as the number of neighboring FA the packets are multicast is fewer than the two other cases.

It is also to note that the handoff is not successful for WLAN→WLAN and WLAN→GPRS at speeds for 22, 25, 28 and 30 m/s. This is due to the fact that the MH is traveling the whole distance from one end of the small WLAN cell to the other before handoff can take place for those high speeds and the threshold time (10 seconds). This is another problem for the handoff algorithm based on threshold time.

4.6.3.1. Proposed Improvement in Overhead

In this section, we propose an algorithm to improve the overhead in multicasting handoff process. The idea for our algorithm is that if we can adapt the Threshold Time dynamically with the increase or decrease of the speed, we can solve the overhead problem. Here we introduce Threshold Distance (D), a parameter on which the handoff decision will take place. The implementation of the algorithm is schematically shown in Figure 4.22.

We define the Threshold Distance (D) to be the product of Threshold Time (T_{th}) and the speed (V).

$$D = V \times T_{th} \qquad (4.11)$$

The Distance Threshold values are calculated from Absolute Received Signal Strength (ARSS) and Relative Received Signal Strength (RRSS). In our simulation, we have considered the following Hysteresis (described in section 3.2) margins:

For ARSS: 3 dB

For RRSS: 6 db

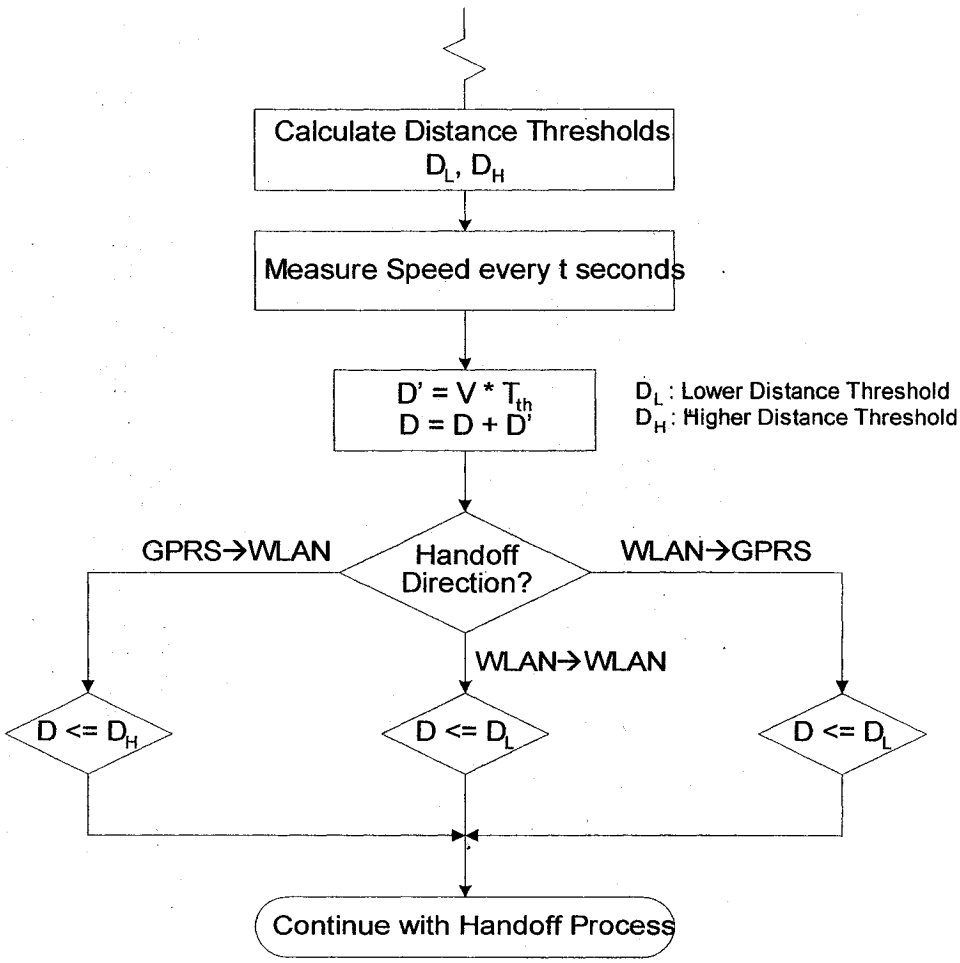

Figure 4.22: Algorithm for Improved TOHR

The algorithm uses both the signal strength measurements to find the two threshold values. The lower value, denoted as D_L, is the threshold point for WLAN→WLAN and WLAN→GPRS handoffs. On the other hand, the higher value, denoted by D_H, is the threshold for GPRS→WLAN handoff. This will ensure better reliability for voice calls for which multicasting scheme is being used.

4.6.3.2. Simulation Results

After implementing the proposed improvement in our algorithm, we have simulated handoff for GPRS→WLAN, WLAN→WLAN and WLAN→GPRS, and studied the Traffic Overhead Ratio (TOHR) at different MH speed. Figure 4.23 shows the change of average TOHR for handoffs simulated with the following parameters:

Speed: 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 30 m/s

Number of handoffs simulated: 10 (in each speed) for GPRS→WLAN, 3 (in each speed) for WLAN→WLAN and 3 (in each speed) for WLAN→GPRS

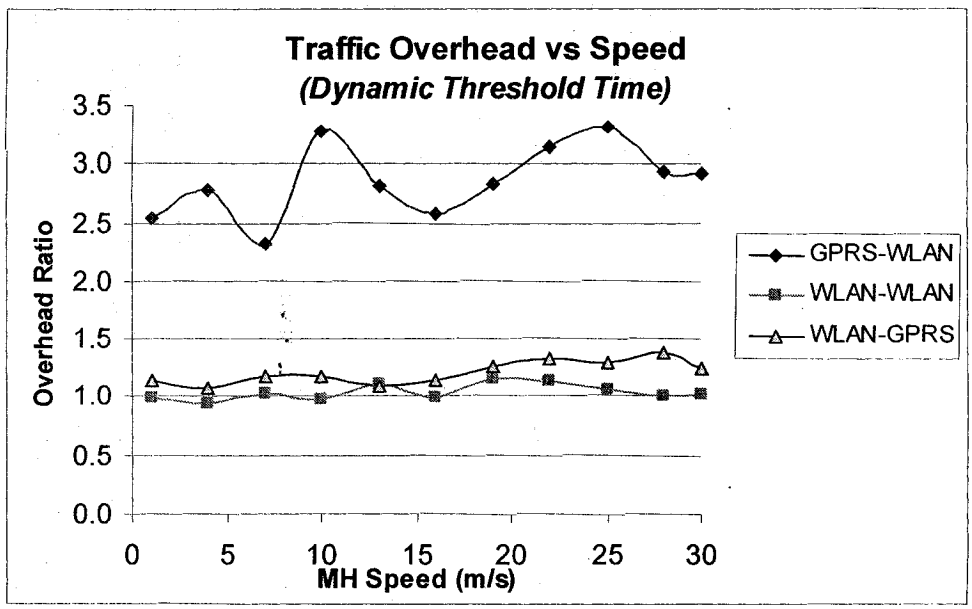

Figure 4.23: Traffic Overhead for various handoff scenario at varying speed with dynamic Threshold Time

From the figure it is evident that the traffic overhead for all the three handoff cases does not change much as the speed of the mobile host increases. Our result shows that the overhead is slightly higher in WLAN→GPRS than in WLAN→WLAN. However, the

overhead is higher for GPRS→WLAN in view of the fact that we choose a higher value for Distance Threshold. The average values of the TOHR are shown in Table 4.4.

Handoff Direction	*Traffic Overhead Ratio (TOHR)*
GPRS→WLAN	2.86
WLAN→WLAN	1.04
WLAN→GPRS	1.21

Table 4.4: Average TOHR using Adaptive Threshold Time Algorithm

Figure 4.24, 4.25 and 4.26 show the Traffic Overhead along with the Threshold Time for various speeds of the mobile host for the handoff cases GPRS→WLAN, WLAN→WLAN and WLAN→GPRS respectively. In these scenarios, the Threshold Time decreases exponentially with the increase of the MH speed.

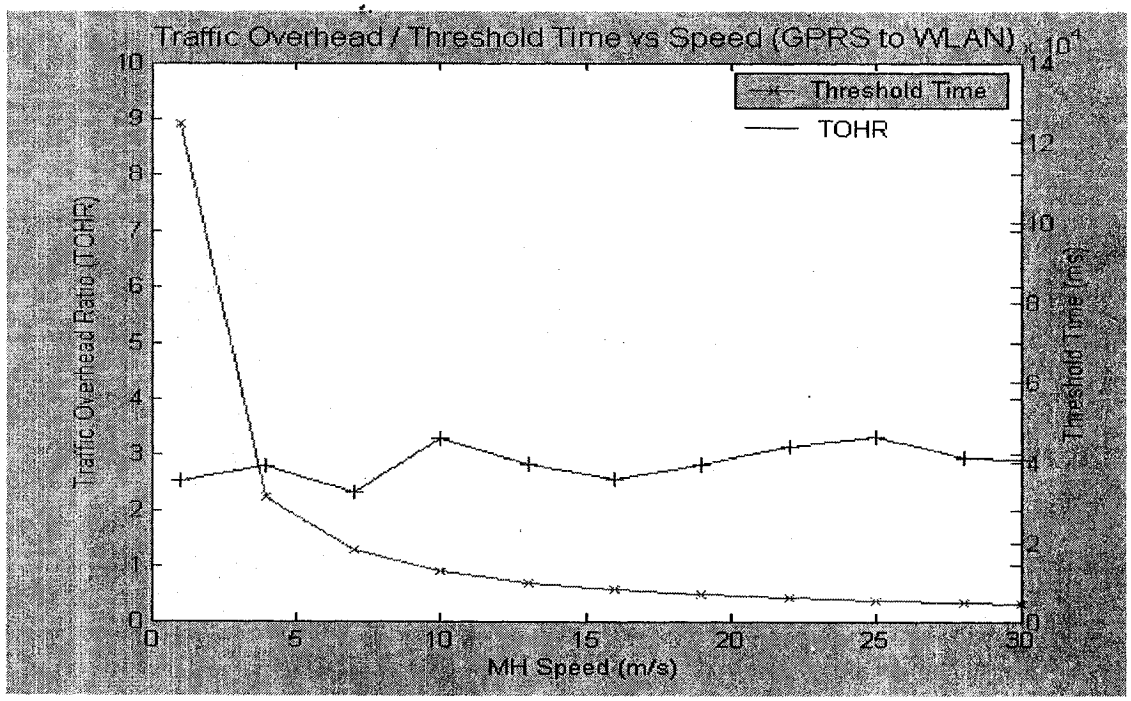

Figure 4.24: Traffic Overhead and Threshold Time at various Speeds (GPRS → WLAN)

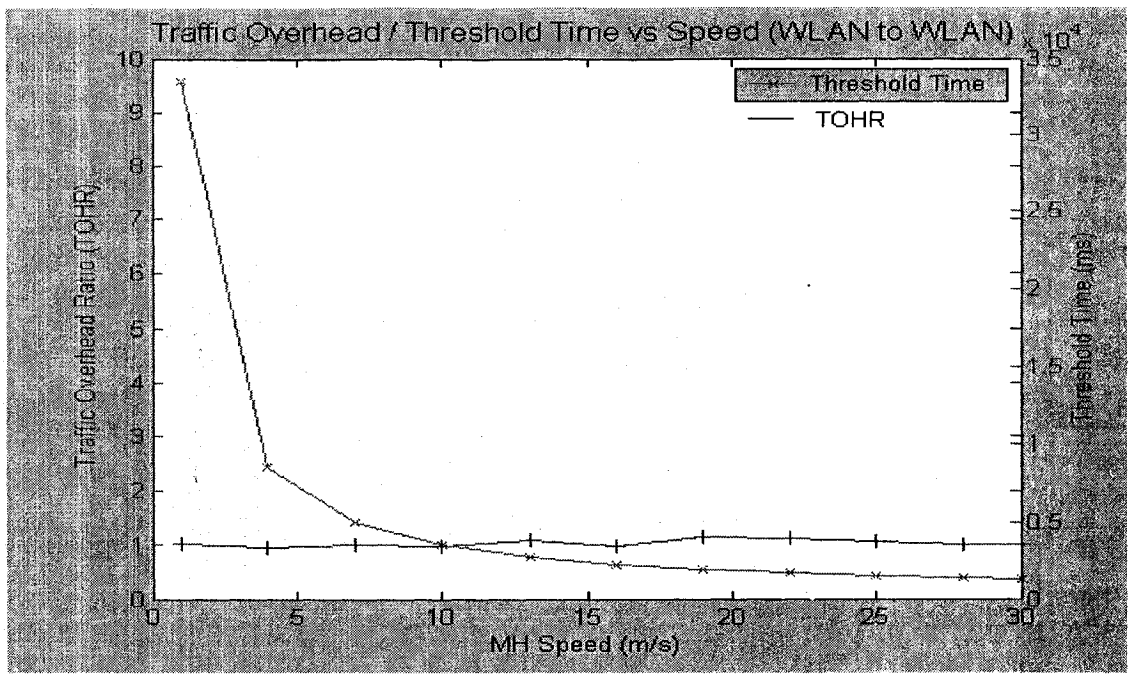

Figure 4.25: Traffic Overhead and Threshold Time at various Speeds (WLAN→WLAN)

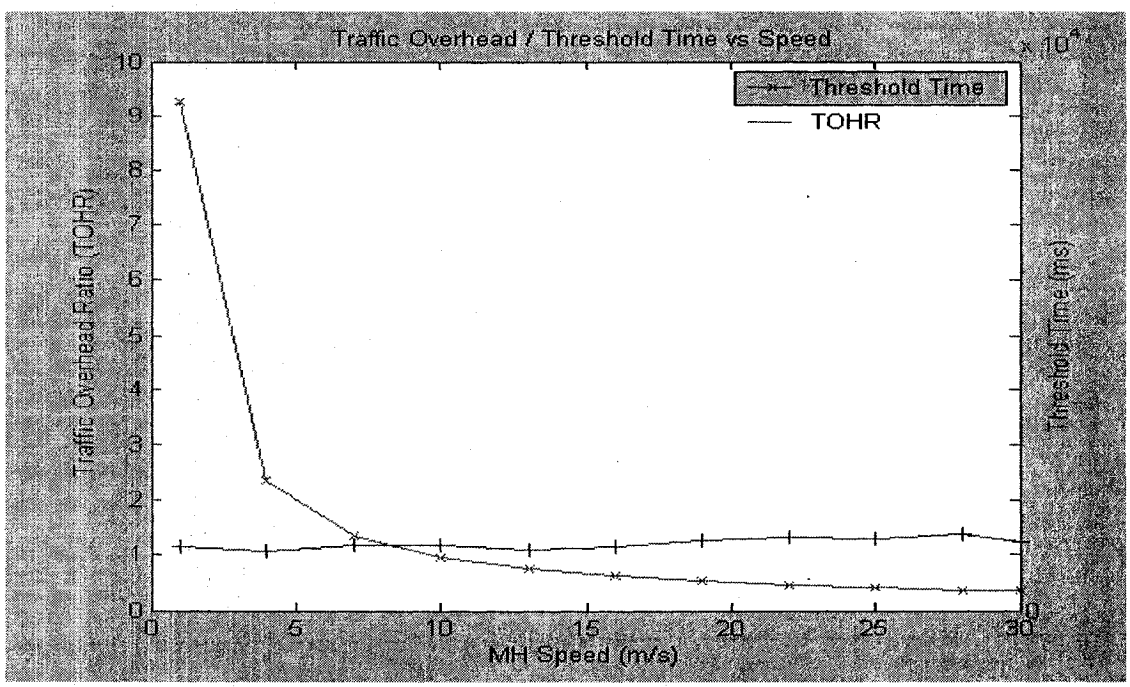

Figure 4.26: Traffic Overhead and Threshold Time at various Speeds (WLAN→GPRS)

4.6.4. Throughput Performance

Our proposed handoff algorithm selects different Distance Threshold values based on the following two criteria:

(1) Call Direction: GPRS → WLAN, WLAN → WLAN, and WLAN → GPRS

(2) Call Type: Voice Call, and Data Call

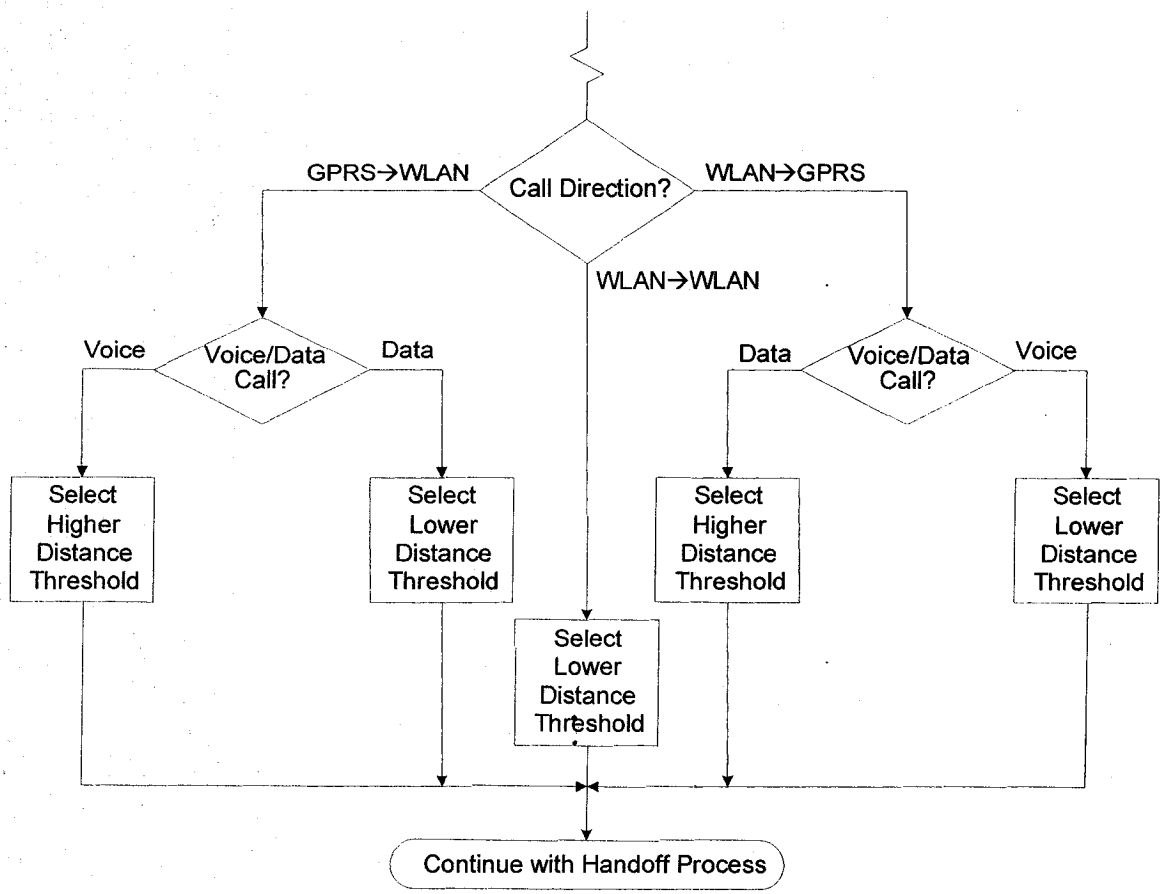

Figure 4.27: Selection of Distance Threshold

Figure 4.27 illustrates the selection process of the Threshold Distances. There are two Distance Threshold values to be calculated in the algorithm based on the following two criteria:

(1) Signal Strength Measurements – ARSS and RRSS

(2) Assumption of the Hysteresis values for ARSS and RRSS cases

4.6.4.1. Proposed Improvement in Throughput

The above algorithm will improve two important Quality of Service (QoS) criteria in the handoff process as described below:

a. ***Improved Throughput:*** The throughput will improve as the data calls, which are normally of much higher data rates than those of voice call, assume higher Distance Thresholds. The higher the Distance Threshold the larger the handoff transition period. As a result, high data rate calls will stay with the WLAN network to take advantage of the large channel capacity of the latter. On the other hand, the quicker handover of the voice calls, which normally take lower bandwidth, to the GPRS cell will not change the throughput. As a result, the overall throughput will increase. Besides, the data calls from the GPRS network, which has a limited channel capacity, will be handed off to the WLAN cell quickly to take advantage of the large channel capacity.

b. ***Better Reliability:*** The reliability will improve as the GPRS network, which is much reliable to handle voice services; will serve the voice calls for a longer period of time through adoption of higher Distance Threshold. Besides, the voice calls at the WLAN network are quickly handed off to the GPRS network. Thus a better reliability is achieved.

4.6.4.2. Simulation Results

We have simulated handoffs using our proposed throughput improvement algorithm for various probabilities of data calls. Our simulation provides throughput (in Kb) during the handoff transition time for two Distance Threshold for each probability of call type distributions – (i) proposed threshold, and (ii) average threshold (i.e. average of the higher and lower values of the distance thresholds).

Figure 4.28 (WLAN→GPRS handoffs only) and Figure 4.29 (combining WLAN→GPRS and GPRS→WLAN handoffs) show the average throughput during handoff transition time. For the second case (i.e. Figure 4.29) we have considered equal proportion of handoffs in both the directions).

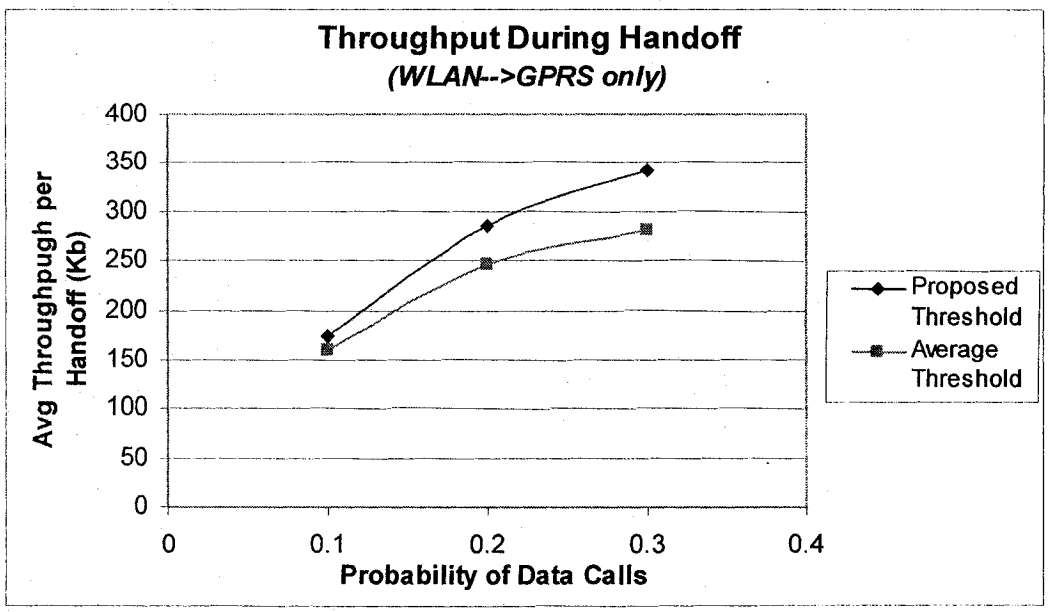

Figure 4.28: Comparison of average throughput during the transition time for WLAN→GPRS handoff with varying probability of data calls.

The graph has been plotted for various probabilities of data calls. The remaining probability is that of voice calls (i.e. if probability of data call is 0.2, then the probability of voice call is 0.8). We have taken the average simulated values of 30 handoffs for each measurement of probability of data calls (as for samples of size n >= 30, regardless of shape of most populations, sampling theory guarantees good results). The figures also show the comparison of throughput between use of the proposed threshold and the average threshold.

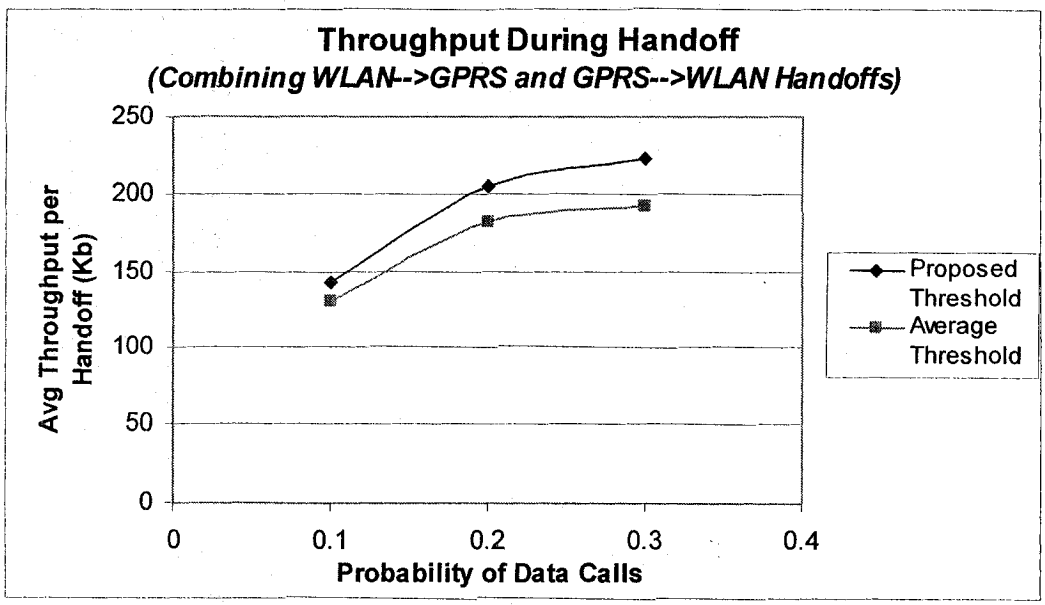

Figure 4.29: Comparison of average throughput for WLAN→GPRS and GPRS→WLAN handoff with varying probability of data calls.

As we observe the results depicted in the graph, the average throughputs per handoff during handoff transition time for the proposed threshold values are higher than those for the average threshold values in all the measured cases. Thus it is evident that our proposed threshold selection algorithm provides better results in terms of throughput. The

graph also shows that the throughput during handoff increases with the increase of the probability of data calls. It is also evident that our algorithm will produce better result at higher proportion of data calls.

We have also simulated handoffs using our proposed throughput improvement algorithm for various cell sizes. Our simulation provides throughput (in Kb) during the handoff transition time for various probabilities of voice calls (70%, 80%, 90%).

Figure 4.30 (GPRS→WLAN), Figure 4.31 (WLAN→GPRS) and Figure 4.32 (combining WLAN→GPRS and GPRS→WLAN) show the average throughput during handoff transition time. For the third case (i.e. Figure 4.32) we have considered equal proportion of handoffs in both the directions).

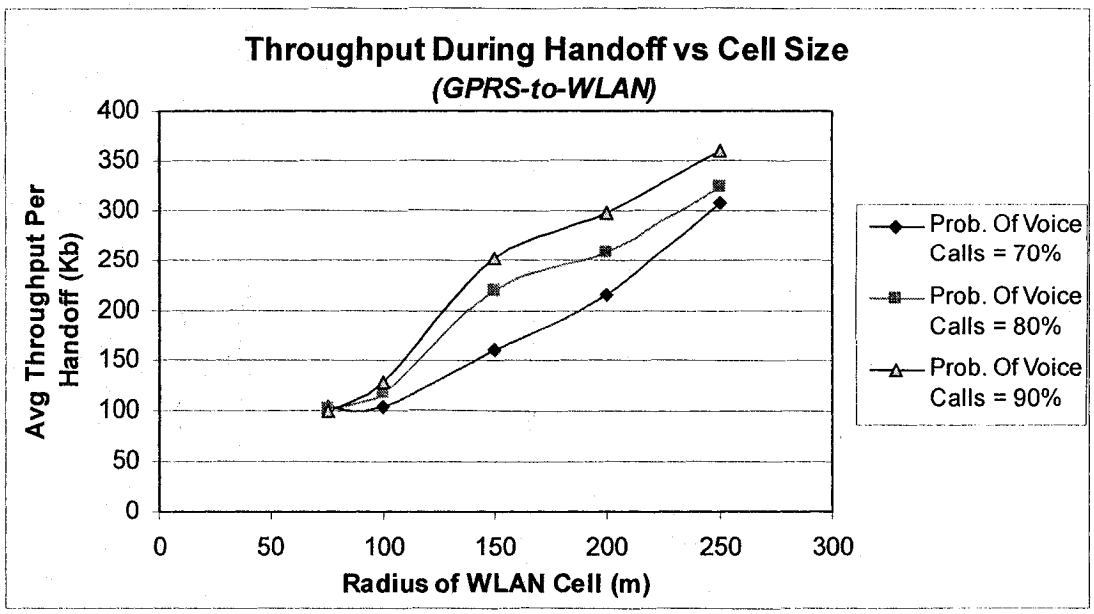

Figure 4.30: Average throughput vs. cell area during handoff for GPRS→WLAN handoffs at varying probability of voice calls

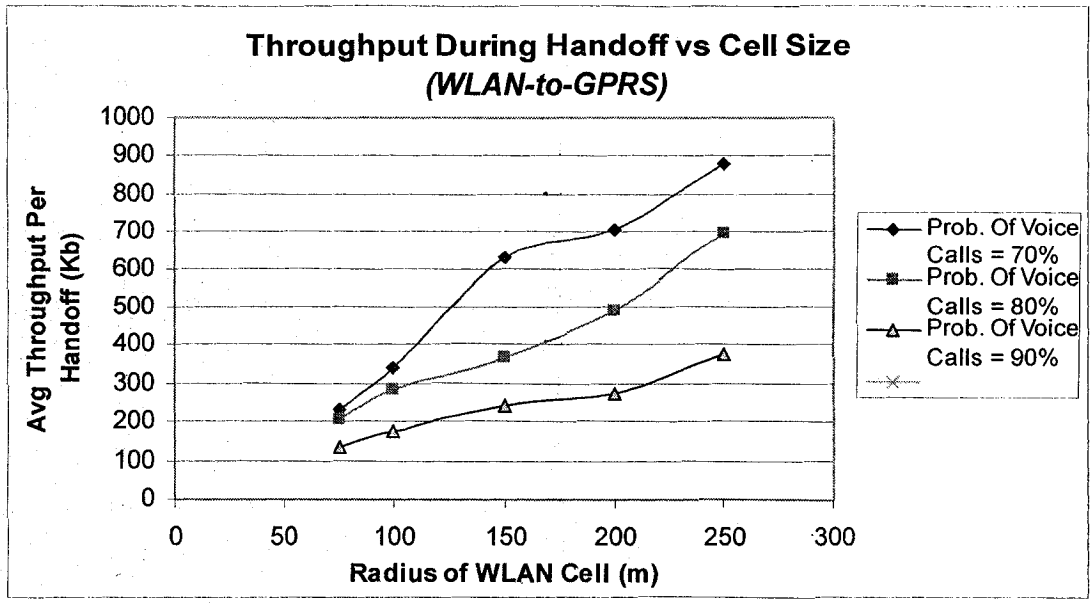

Figure 4.31: Average throughput vs. cell area during handoff for WLAN→GPRS handoffs at varying probability of voice calls

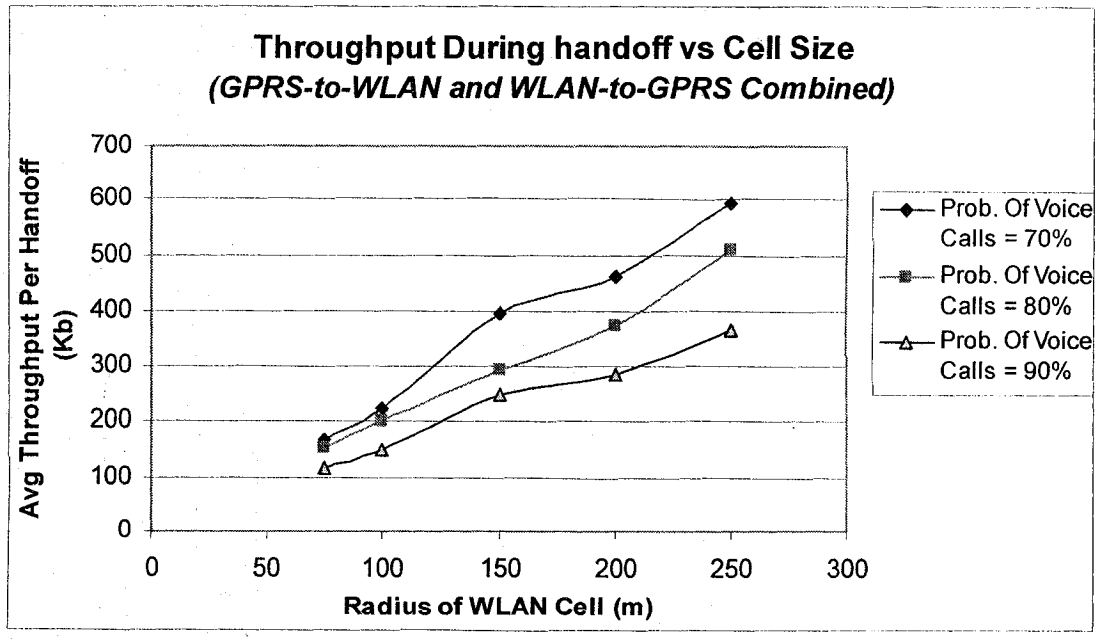

Figure 4.32: Average throughput vs. cell area during handoff for averaging GPRS→WLAN and WLAN→GPRS handoffs at varying probability of voice calls

It is evident from the graphs that the average throughput during handoff transition period increase with the increase of cell radius (in our simulation, we varied the radius of the WLAN cells for the values 75, 100, 150, 200 and 250 meters) for both the call directions (GPRS→WLAN and WLAN→GPRS). It is noticeable that for GPRS→WLAN handoffs, the throughput is higher for higher probabilities of voice calls. On the other hand, for WLAN→GPRS the throughput is higher for lower probabilities of voice calls (i.e. higher probabilities of data calls). However, the combined results show that the throughput increases for decrease of probability of voice calls (increase of probability of data calls). This phenomenon is due to longer handoff transition time for the MH within the coverage of the WLAN cell for data calls.

4.6.5. Call Drop

Call Drop depends on several factor; e.g. for low signal level, unavailability of traffic channels in the target AP/BS etc. In this thesis, we propose the use of Hysteresis margin to improve the percentage of drop call due to weak signal. Drop Call due to signal quality is likely to decrease with the decrease of Hysteresis margin. However, we have used only one set of Hysteresis margin (3 dB and 6 dB for ARSS and RRSS respectively) throughout our simulations.

While we have not simulated call drops for various signal quality, in this section we have observe the scenarios for various channel capacity of the target AP/BS. We have measured the probability of dropped calls for different capacity of AP/BS dedicated for handoff traffic in our simulation. Figure 4.33 and Figure 4.34 shows the results for

WLAN and GPRS cells respectively. We have also obtained results for different values of maximum rate of each call.

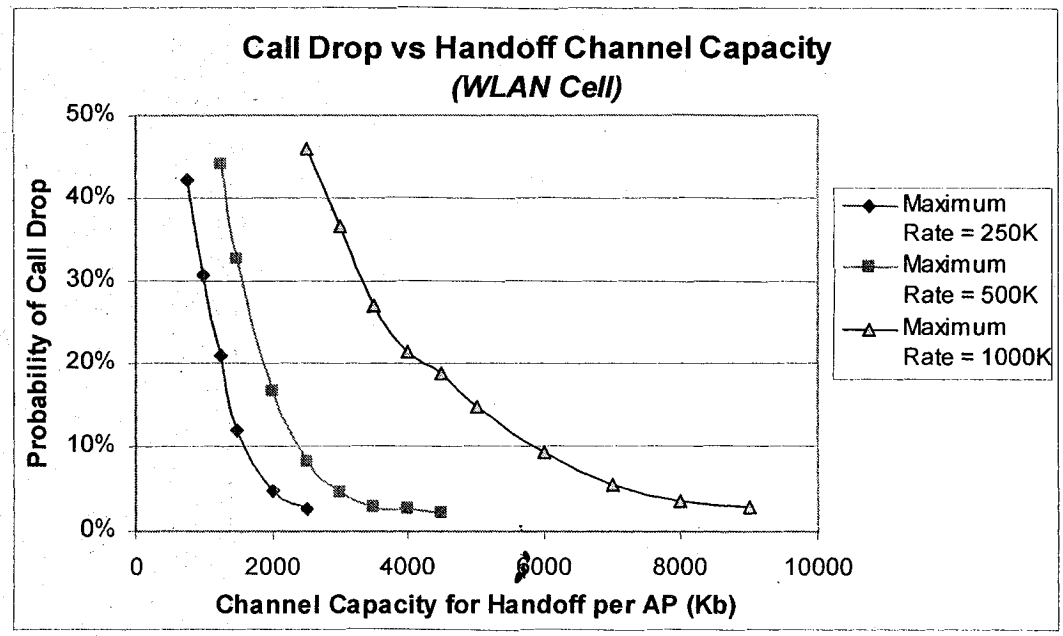

Figure 4.33: Probability of dropped calls with change in channel capacity for handoff for WLAN cells

As we observe the WLAN cell results in Figure 4.32, the probability of dropped calls decreases with the increase of channel capacity for handoff traffic. In the lower range of channel capacity, the probability changes sharply. This phenomenon is due to the fact that the low handoff capacity has more probability to be occupied due to the random nature of data rates at a wide range of 8~1000 Kbps. The change of probability decreases at a higher range of channel capacity. It is also noticeable that the rate of change of probability of call drop is less for higher values of maximum data rate.

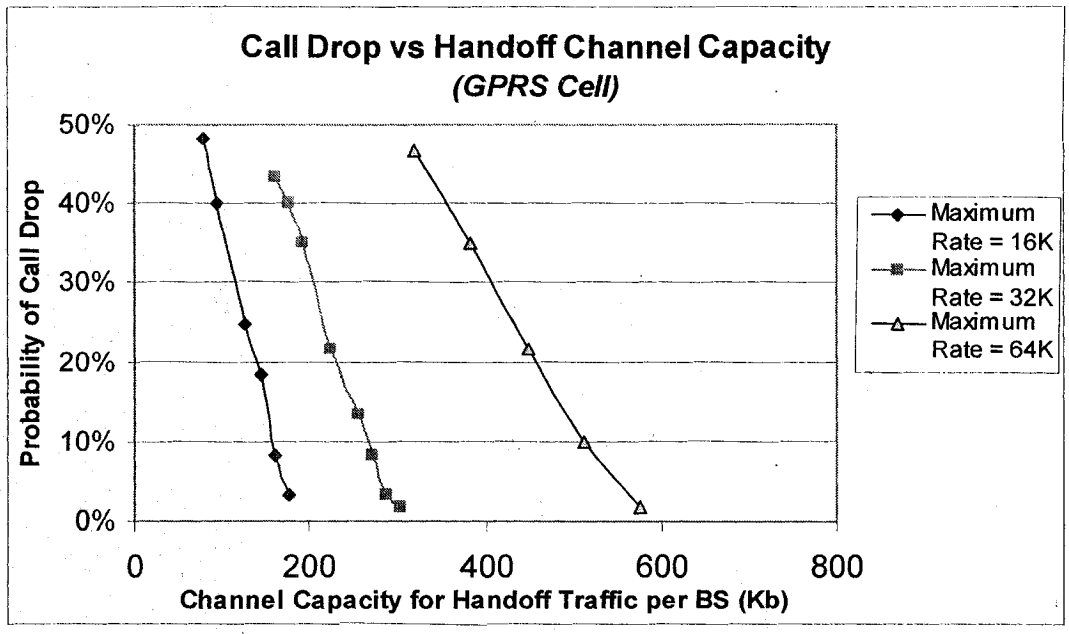

Figure 4.34: Probability of dropped calls with change in channel capacity for handoff for GPRS cells

As shown in Figure 4.33, for GPRS cell, the probability of dropped calls decreases with the increase of channel capacity for handoff traffic as in the case of WLAN cell. The change of the rate of change of probability of call drop is more or less linear for all the three cases of maximum data rate we have studied. However, the steepness of the curves signifies that the probability of drop call decreases at a lesser rate at higher values of maximum data rate.

Chapter 5

5. Conclusions and Future Works

5.1. Conclusions

This thesis presents a new technique for handoff between WLAN and GPRS networks. Simulation results for handoff latency, overhead and throughput have been presented. We summarize our contributions and findings as follows:

1. We have chosen Mobile IP-based mobility architecture for managing handoff between WLAN and GPRS networks. Being the logical evolution of the widely used as well as popular Internet Protocol (IP) technology, Mobile IP-based architecture will be the easiest one to implement.

2. To improve the reliability of the channel quality measurements, we introduce two different signal strength measurements – Absolute Received Signal Strengths (ARSS) and Relative Received Signal Strengths (RRSS). Handoff is implemented if the ARSS or/and RRSS reaches predefined threshold value(s) based on the handoff direction (GPRS→WLAN, WLAN→WLAN or WLAN→GPRS).

3. We have simulated handoffs using the standard Mobile IP signals adopting both Unicasting and Multicasting packet recovery schemes. The results show a significant reduction of handoff latency if Multicasting scheme is adopted. We

propose Multicasting for the voice calls while Unicasting for the data calls, in which delay is not a major problem.

4. We propose two modifications of the Mobile IP signal flows for further improvement of latency using Multicasting. Firstly, the new FA will not wait for the registration process to complete, and start sending packets to the MH after a fixed delay. Secondly, the packet size will be dynamically changing with the change of data rates. We have simulated handoffs using the standard Mobile IP and our proposed algorithm. The results show improvement in handoff latency using our proposed algorithm, particularly at lower data rates at which most of the voice calls operate.

5. While Multicasting improves the latency, it also involves wastage of bandwidth, defined as Traffic Overhead, as data packets are being stored in neighboring FAs. To improve overall traffic overhead, we introduce a new approach in handoff decision making process. We define a new threshold type, namely Distance Threshold, which assumes two values – lower and higher - based on the values of ARSS and RRSS. However, the selection of the Distance Threshold value (lower or higher) will depend of the handoff direction as well as call type (voice or data). The MH will implement handoff as soon as the measured distance it travels in the handoff transition time goes below the appropriate Distance Threshold. Our simulation results show that the traffic overhead ratio can be kept more or less constant, using our proposed algorithm, with the change of the speed.

6. In our thesis, we describe the selection process of the Distance Threshold. For a voice call, the MH will assume a lower threshold value as it is handing off from WLAN to GPRS cell, while the threshold takes a higher value for the opposite direction. For a data calls the selection of values are reversed. Our simulation results shows better throughput using our proposed method. We have also simulated handoffs for various probabilities of voice/data calls. Our results show that the throughput during handoff is improved as for higher probability of data calls.

While our simulation results prove the effectiveness of the scheme, the proposed handoff scheme may lead to some adverse QoS issues. The scheme involves additional signaling which requires additional bandwidth as well as complexity in the network. We have used a simplified model for our simulation. The results may differ in the scenario of real life networks. Furthermore, there are some other QoS parameters, except the three we have studied, which need to be investigated. Nevertheless, our proposed handoff scheme may be considered as a base model for future heterogeneous networks which can be further improved.

5.2. Future Works

We suggest adding the following works for future research:

1. We have simplified our analysis as we have considered distance as the only factor for degradation of signal strength. Other factors, such as fading, interference, noise etc., may be considered in future studies.

2. While we have defined the mobility architecture as overlay-underlay networks, we have not considered this effect in our simulation. In future studies, fast users may be encouraged to join GPRS (overlay network), and slow users to join WLAN (underlay network). This will jointly reduce the number of handoffs and increases the total system capacity.

3. In describing Multicasting scheme, we have considered the FA to send data packets all the neighboring FAs. If the MH knows which AP/BS it is moving towards, the traffic overhead can be reduced significantly. The detection of movement pattern can be incorporated in our handoff technique in future studies.

4. We have only considered either voice (real time) or data (non-real time) services in our study. However, multimedia applications, which comprise both voice and data, may be considered in future studies.

5. We have only studied the handoff performance related to the three QoS parameters – latency, overhead, and throughput. However, there are some other parameters (e.g. Call Drop, Call Blocking etc.) to be worked on for further improvement.

Appendix I

Bibliography

[1] Apostolis K. Salkintzis, Chad Fors, and Rajesh Pazhyannur, "WLAN-GPRS Integration for Next-Generation Mobile Data Networks", IEEE Wireless Communications, October 2002, pp 112-124

[2] Malathi Veeraraghavan, Nabeel Cocker, and Tim Moors, "Support of Voice Services in IEEE 802.11 Wireless LANs", IEEE Infocom 2001, pp 488-497

[3] Sumi Helal, Choonhwa Lee, Yongguang Zhang, Golden G. Richard III, "An Architecture for Wireless LAN/WAN Integration", Wireless Communications and Networking Conference, 2000. WCNC. 2000 IEEE ,Volume: 3 , 23-28 Sept. 2000 p1035-1041

[4] Mark Stemm, and Randy H. Katz, "Vertical handoffs in Wireless Overlay Networks", Baltzer Science Publishers BV, Mobile Network and Applications,, Volume 3, Issue 4, 1998, pp 335-350

[5] Min-hua Ye, Yu Liu, Hui-min Zhang, "Mobile IP Handoff Between Hybrid Networks", PIMRC (Personal, Indoor and Mobile Radio Communications) 2002, 13th IEEE International Symposium, Sept. 2002, Vol. 1, pp 265-269

[6] Wei Wu, Wen-Shiung Chen, Ho-En Liao, and Fongray Frank Young, "A Seamless Handoff Approach of Mobile IP Protocol for Mobile Wireless Networks", IEEE Transactions on Consumer Electronics, Vol. 48, Issue 2, May 2002, pp 335-344

[7] Robert Berezdivin, Robert Breinig, and Randy Topp, "Next-Generation Wireless Communications Concepts and Technologies", IEEE Communications Magazines, March 2002, Vol. 30, Issue 3, pp 108-116

[8] Prasan de Silva, and Harsha Sirisena, "A Mobility Management Protocol for IP-Based Cellular Networks", IEEE Wireless Communication, Vol. 9, Issue 3, June 2002, pp 476-482

[9] Neeli Prasad, Anand Prasad, editors, "WLAN Systems and Wireless IP for Next Generation Communications", Artech House Publishers, Boston and London, 2002

[10] Jim Geier, "Wireless LANs: Implementing Interoperable Networks", Macmillan Technical Publishing, First Edition, 2001

[11] Code of Federal Regulation, Federal Communications Commission, Part 15, Section 15.247, pp 740-742, Revised as of October, 2002, Online Link: http://a257.g.akamaitech.net/7/257/2422/14mar20010800/edocket.access.gpo.gov/cfr_2002/octqtr/47cfr15.247.htm

[12] Alberto Leon-Garcia, Indra Widjaja, "Communication Networks: Fundamental Concepts and Key Architecture", McGraw-Hill, 2002

[13] Behrouz A. Forouzan, "Local Area Networks", First Edition, McGraw-Hill, New York, 2003

[14] William Stallings, "Wireless Communications and Networks", Prentice Hall, New Jersey, 2002

[15] Vijay K. Garg, "Wireless Network Evolution: 2G to 3G", Prentice Hall, NJ, 2001

[16] Emmanuael Seurre, Patrick Savelli, and Pierre-Jean Pietri, "GPRS for Mobile Internet", Archtech House, Norwood, MA, 2003

[17] Abbas Jamalipour, "The Wireless Mobile Internet: Architecture, Protocols and Services", John Wiley & Sons, England, 2003

[18] Jian Cai, David J. Goodman, "General Packet Radio Service in GSM", IEEE Communication Magazine, Vol. 35, Issue 10, pp 122-131, October 1997

[19] J. Parantainen, S. Hatimi, "Delay Analysis for IP Speech Over GPRS", IEEE Vehicular Technology Conference 1999, Volume 2, pp 829-833, 19-22 September, 1999

[20] Nikhil M. Deshpande, Jay Gilbert, "GPRS – How Does It Work and How Good Is It", Intel Developer Update Magazine, October 2002, Online Link: http://www.intel.com/update/departments/wireless/wi10021.pdf

[21] Intel Corporation, White Paper, 2002, Online Link: http://www.techonline.com/pdf/pavillions/intel/gprs.pdf

[22] A. Lakaniemi, J. Parantainen, "On Voice Quality of IP Voice Over GPRS", IEEE International Conference on Multimedia and Expo, 2000, Volume 2, pp 751-754, July 30 - Aug. 2, 2000

[23] GPRS White Paper, Cisco, July, 2002, Online Link: http://www.cisco.com/warp/public/cc/so/neso/gprs/gprs_wp.htm

[24] Charles C. Perkins, Tutorial titled "Mobile Networking Through Mobile IP", Online Link: http://www.computer.org/internet/v2n1/perkins.htm

[25] "IP Mobility Support," Charles. Perkins, ed., IETF RFC 2002, Oct. 1996, Online Link: http://www.ietf.org/rfc/rfc2002.txt

[26] J. B. Postel, ed.,"Internet Protocol", RFC 791, September 1981

[27] Debalina Ghosh, "Mobile IP: Connecting the World", ACM Crossroads Student Magazine, January 2001

[28] Charles C. Perkins, "Mobile IP", IEEE Communications Magazine, May 1997, pp 84-99

[29] S.E. Deering, ed., "ICMP Router Discovery Messages",IETF RFC 1256, Sept. 1991, Online Link: http://www.ietf.org/rfc/rfc1256.txt

[30] Christna A Nika, Dimitrios D. Vergados, and Michael Theologou, "Mobile IP: A Challenge in the Mobile World", the fourth International Symposium on Wireless Personal Multimedia Communications, September 9-12, 2001, Aalborg, Denmark

[31] Charles C. Perkins, "Mobile IP: Design Principles and Practices", Reading, MA, Addision-Wesley, 1998

[32] R.L. Rivest, "The MD5 Message-Digest Algorithm," IETF RFC 1321, Apr. 1992, Online Link: http://www.ietf.org/rfc/rfc1321.txt

[33] Charles C. Perkins, ed., "IP Encapsulation within IP", RFC 2003, October 1996, Online Link: http://rfc.sunsite.dk/rfc/rfc2003.html

[34] David B. Johnson and Charles E. Perkins, "Route Optimization in Mobile IP", Internet Draft, February 1999

[35] Charles C. Perkins, ed., "Minimal Encapsulation within IP", IETF RFC 2004, October 1996, Online Link: http://www.ietf.org/rfc/rfc2004.txt

[36] Theodore S. Rappaport, "Wireless Communications: Principles and Practice", second edition, 2002, Prentice Hall, Upper Sadle River, NJ 07458

[37] Mikael Gudmundson, "Analysis of Handover Algorithms.", IEEE Vehicular Technology Conference, pp537-542, July 1991.

[38] Gregory P. Pollini, "Trends in Handover Design", IEEE Communications Magazine, March 1996, pp 82-90

[39] Kaven Pahlavan, Prashant Krishnamurthy, Ahmad Hatami, et al, "Handoff in Hybrid Mobile Data Network", IEEE Personal Communications, pp 34-47, April 2000

[40] Ahmed Helmy, "A Multicast-based Protocol for IP Mobility Support", ACM Press, New York, NY, pp 49-58, 2000

[41] Eunsoo Shim, Hung-yu Wei, Yusun Chang, and Richard D. Gitlin, "Low Latency Handoff for Wireless IP QOS with Neighborcasting", IEEE International Conference on Communications 2002, Vol. 5, pp 3245-3249, April 28 – May 2, 2002

[42] Gary C. Kessler, "Future Mobility: Mobile IP is the Harbinger of Untethered Computing", Telephony Online (Primedia Publication), September 1998, Online Link: http://telephonyonline.com/ar/telecom_future_mobility_mobile/

[43] Apostolis K. Salkintzis, "Interworking Between WLANs and Third-Generation Cellular Data Networks", IEEE Vehicular Technology Conference 2003, Vol. 3, pp 1802-1806, April 22-25, 2003

[44] Hyosoon Park, Sunghoon Yoon, Taehyoun Kim, Jungshin Park, Misun Do, and Jaiyoung Lee, "Vertical Handoff Procedure and Algorithm Between IEEE 802.11 WLAN and CDMA Cellular Network", LNCS2524, pp 103-112, 2003

[45] Ylianttila M., Pande M., Maleka M., Mahonen P., "Optimization Scheme for Mobile Users Performing Vertical Handoffs Between IEEE 802.11 and GPRS/EDGE Networks", IEEE GLOBECOM '01, Vol. 6, pp 3439-3443, Nov. 25-29, 2001

[46] Min-hua Ye, Zhe-wei Wang, Hui-min Zhang, and Yu Liu, "Performance Analysis of TCP/UDP During Mobile IP Handoffs", International Conferences on Info-tech and Info net, Beijing. 2001, Vol. 2, pp 724-729, 2001.

[47] Antoine Stephane, and A. H. Aghvami, "Fast handover Schemes for Future Wireless IP Networks: a Proposal and Analysis", Vehicular Technology Conference 2001, Vol. 2, pp 605-609, 7-11 Oct. 2001

[48] "ETSI HIPERLAN/2 Standard", Online link:
http://portal.etsi.org/bran/kta/Hiperlan/hiperlan2.asp

[49] Kuo-Hsing Chiang, Nirmala Shenoy, "A Random Walk Mobility Model for Location Management in Wireless Networks", IEEE PIMRC, 2001

[50] Seshadri Mohan, and Ravi Jain, "Two User Location Strategies for Personal Communications Services", IEEE Personal Communication, Vol. 1, Issue 1, pp 42-50, 1st Qtr 1994

[51] Kathleen S. Meier-Hellstern, and Eduardo Alonso, "The Use of SS7 and GSM to Support High Density Personal Communications", IEEE SUPERCOM/ICC 92, Vol. 3, pp 1698-1702, 14-18 June, 1992

[52] Ian F. Akyildiz, and Weyne Wang, "A Dynamic Location management Scheme for Next-Generation Multitier PCS Systems", IEEE Transactions of Wireless Communications, Vol. 1, No. 1, January 2002

[53] Website of Nokia, Online link:
http://www.nokia.com/nokia/0,1522,,00.html?orig=/gprs

[54] Pyramid Research, "The Great Standard Migration: IS-136 TDMA Moves to GSM/GPRS Overlay", July 2001, Online link:
http://www.pyr.com/info/rpts/july01_latdma.asp

Appendix II

Glossary

1G: First Generation

2G: Second Generation

3G: Third Generation

AP: Access Point

BS: Base Station

BSC: Base Station Controller

BSS: Basic Service Set

CH: Correspondent Host

CoA: Care-of Address

CP: Contention Period

CFP: Contention-free Period

CS: Coding Scheme

CSMA-CA: Carrier Sense Multiple Access with Collision Avoidance

DCF: Distributed Coordination Function

DIFS: DCF Interframe Space

DS: Distribution System

DSSS: Direct Sequence Spread Spectrum

ESS: Extended Service Set

ETSI: European Telecommunications Standards Institute

FA: Foreign Agent

FCC: Federal Communications Commission

FHSS: Frequency Hopping Spread Spectrum

GGSN: Gateway GPRS Support Node

GSM: Global System for Mobile Communications

GPRS: General Packet Radio Service

GSN: GPRS Support Node

HA: Home Agent

HLR: Home Location Register

HO: Handoff

HYPERLAN: High Performance Radio Local Area Network

IEEE: Institute of Electrical and Electronic Engineers

IETF: Internet Engineering Task Force

IFS: Interframe Space

IP: Internet Protocol

IPv4: Internet Protocol version 4

IPv6: Internet Protocol version 6

ISM: Industrial, Scientific, and Medical

LAN: Local Area Network

LLC: Logical Link Control

MAC: Medium Access Control

MD: Message Digest

MG: Mobility Gateway

MH: Mobile Host

MS: Mobile Station

NAV: Network Allocation Vector

OSI: Open System Interconnection

PCF: Point Coordination Function

PCU: Packet Control Unit

PDCH: Packet Data Channel

PDN: Packet Data Network

PHY: Physical Layer

PIFS: PCF Interframe Space

QoS: Quality of Service

RA: Routing Area

RFC: Request for Comment

RLC: Radio Link Control

RSS: Received Signal Strength

RTP: Real Time Protocol

SAP: Service Access Points

SDU: Service Data Unit

SIFS: Short Interframe Space

SGSN: Serving GPRS Support Node

TCP: Transmission Control Protocol

UDP: User Datagram Protocol

WLAN: Wireless Local Area Network